State-of-the-Art

Compressieformaten – Metadatastandaarden – Containerformaten

BOM-Vlaanderen

WP3: Metadatastandaarden en Uitwisselingsformaten

Deliverable 1 (D.3.1)

30 september 2008

UGent – MMLab: Erik Mannens, Sam Coppens, Brecht Vekeman, Rik Van de walle

UGent – Bibliotheek: Patrick Hochstenbach, Paul Bastijns, Siska Corneillie, Liesbeth Van Melle

Inhoud

1 Inleiding en probleemstelling

De preservatie van digitale multimedia-informatie stelt bijzondere eisen aan archiefomgevingen. Enerzijds moeten software- en hardwareoplossingen de toegang tot informatie gedurende lange tijd garanderen. Anderzijds staat ook menselijke input, in de vorm van archiefbeschrijvingen, werkprocessen en het gebruik van standaarden, er voor in dat informatie zo lang mogelijk beschikbaar én interpreteerbaar blijft voor grote gebruikersgroepen.

Digitale informatie staat immers aan veel gevaren bloot. Sommige daarvan bedreigen ook analoge documenten, andere zijn enkel kenmerkend voor digitale informatie:

- In digitale vorm is informatie een conceptueel object. Digitale multimedia kunnen gemakkelijk gekopieerd en gewijzigd worden zonder onmiddellijk zichtbare effecten op de representeerbare inhoud. In vergelijking met analoge informatie is het daarom moeilijker de authenticiteit van informatie in digitale vorm te bewaren. Op korte termijn kunnen hardware, software en menselijke fouten tot verlies van data leiden. Vaak worden deze fouten onmiddellijk opgelost door specialistische correctiemethodes. In andere gevallen worden deze datacorrupties pas in een latere fase opgemerkt, op een moment dat de data al schijnbaar correct verwerkt zijn zonder dat rekening gehouden is met alle technische en visuele aspecten van de intellectuele inhoud.

- Door technologische wijzigingen kunnen dataformaten op termijn in onbruik raken. Ook de levensduur van opslagtechnieken is eindig. Om toegang tot de informatie te garanderen, zijn migratie- of emulatieplannen noodzakelijk. Technische metadata moeten voldoende informatie over de opgeslagen data aanreiken om een tijdig ingrijpen mogelijk te maken.

- Op lange termijn verandert het kennisdomein van gebruikersgroepen, dataspecialisten verdwijnen, organisaties wijzigen of krijgen een nieuwe opdracht. Het gevaar bestaat dat oudere opgeslagen data voor een nieuwe gebruikersgeneratie niet meer interpreteerbaar zijn. De opgeslagen data moeten daarom van voldoende contextuele data voorzien worden opdat deze nieuwe gebruikersgroepen de informatie nog kunnen interpreteren.

Naast de verschillende bestandsformaten waarin multimedia voorkomen, stelt ook hun specifiek toepassingsgebied andere eisen aan de descriptieve metadata. Digitale beeldbestanden in een bibliotheek kunnen een gescand boek representeren dat met bibliografische metadata beschreven moet worden. In een museum zullen beeldbestanden dan weer kunstwerken beschrijven met andere descriptieve metadatavelden. Een videobestand kan bestaan uit een aflevering van een televisie-uitzending, maar het kan ook deel uitmaken van een installatie van een videokunstenaar. In het eerste geval beschrijven de descriptieve metadata de serie en de aflevering waarin de video uitgezonden werd, terwijl in het tweede geval de kunstenaar en de specificaties van de installatie beschreven worden.

Voor de beschrijving van de multimedia zullen voor elke sector specifieke eisen gelden. Maar voor de consultatie van het archief is er behoefte aan een overkoepelend descriptief metadatamodel om zoekacties over de totale dataset mogelijk te maken.

De doelstelling van dit rapport is om verder inzicht te krijgen in:

- de structurele eisen voor de beschrijving van digitale informatie,
- de gangbare bestandsformaten en compressietechnieken voor de opslag en uitwisseling van multimediadata,
- de beschikbare descriptieve metadatastandaarden die in elke sector gebruikt worden,
- de nodige metadata om de authenticiteit van de gedigitaliseerde informatie te bewaren,
- de nodige technische metadata om alle kenmerken van de individuele bestanden te beschrijven,
- de best-practices bij de langetermijnbewaring van multimedia,
- het nodige datamodel voor de langetermijnbewaring van digitale informatie.

In het eerste deel van dit rapport wordt het OAIS-model (ISO-14721) beschreven. OAIS is een conceptueel referentiemodel dat richtlijnen biedt bij de opzet van een digitaal archief met het oog op langetermijnbewaring.

Het tweede deel geeft een overzicht van een aantal standaarden die in het bibliotheekwezen, de omroepsector, de culturele sector, en de erfgoedsector gebruikt worden, in het bijzonder metadatastandaarden en ontologieën (descriptieve, technische, administratieve), container- en compressieformaten.

Vervolgens worden in het derde deel twee praktijkvoorbeelden beschreven: de ontwikkeling van het e-depot in de Koninklijke Bibliotheek van Nederland en de opzet van een Europese meertalige zoekmachine voor cultureel erfgoedonderzoek. De focus in de twee projecten verschilt: in het eerste staat langetermijnbewaring van digitale objecten centraal terwijl het tweede voornamelijk over de ontsluiting en consultatie van het archief gaat.

Ten slotte worden een aantal conclusies geformuleerd in verband met het gelaagd metadatamodel dat in het kader van het BOM Vlaanderen-project zal geconstrueerd worden.

2 Open Archief Informatie Systeem (OAIS)

2.1 Geschiedenis

In 1982 werd het Adviserend Orgaan voor Ruimtevaart Data Systemen (CCSDS) opgericht. Het ging om een internationaal forum van ruimtevaartorganisaties die geïnteresseerd waren in de ontwikkeling van standaarden voor data-uitwisseling ten behoeve van onderzoek. De studies kregen de aandacht van de Internationale Organisatie voor Standaardisatie (ISO), die in 1990 een plan voorstelde om de voorschriften van CCSDS in een standaardisatieprogramma te formaliseren. Op vraag van het ISO technisch comité ISO/TC20/SC13 begon het CCSDS aan een voorstudie over de langetermijnbewaring van digitale databestanden m.b.t. ruimtevaartmissies. Binnen een internationale samenwerking resulteerde dit in 1995 in het conceptueel raamwerk, OAIS genaamd, dat als basis kon dienen voor verdere standaardisatieactiviteiten (CCSDS 2002).

Vanaf de start van het project bleek al dat de OAIS-studie niet enkel relevant was binnen ruimtevaartorganisaties maar ook toepassingen kon vinden in andere uiteenlopende projecten. Zo toonden staatsinstellingen, bedrijven en universiteiten grote belangstelling voor de resultaten. Via workshops en debatten werd het model in een bredere context geplaatst en aangepast. Dit resulteerde in 1997 en 1999 in nieuwe ontwerpen, die in 2000 door het ISO aanvaard werden als een conceptstandaard. Het OAIS-referentiemodel werd in 2002 goedgekeurd als de internationale ISO-standaard 14721. Lavoie (Lavoie 2004) geeft een heldere samenvatting van het OAIS-model.

Anno 2008 worden OAIS-concepten wereldwijd toegepast in digitale archieven. De term “OAIS-compliant” is een handelsmerk geworden voor vele commerciële archieven (zie bijv. IBM’s DIAS, OCLC’s Digital Archive Service, Ex Libris’ DigiTool). Informatiearchitecten in de bibliotheek- en archiefwereld werken aan dataformaten zoals METS en MPEG-21/DIDL, die implementaties zijn van OAIS-informatiepakketten. Metadatastandaarden zoals PREMIS, mede ontwikkeld door bibliotheken, musea, staatsinstellingen en bedrijven, complementeren OAIS op het gebied van preservatiemetadata. Projecten zoals DRAMBORA en TRAC ontwikkelden audit- en certificatiestandaarden voor zogenaamde “trusted digital repositories” in OAIS-stijl.

In WP3 van het project BOM-Vlaanderen werd eveneens geopteerd voor het OAIS-referentiemodel bij de opstelling van een gelaagd metadatamodel dat de preservatie van multimediale data moet garanderen. Er bestaan weliswaar alternatieve referentiemodellen, waaronder CORDRA, DLF en IMS.[1] De bewezen doeltreffendheid en bruikbaarheid binnen genoemde internationale projecten met gelijkaardige doelstellingen als BOM, het grote

[1] Respectievelijk CORDRA (Content Object Repository Discovery and Registration/Resolution Architecture): zie http://cordra.net ; DLF (Digital Library Framework): zie http://www.diglib.org/architectures/serviceframe/ (o.a. Digital Library Federation (2005). DLF service framework for digital libraries : a progress report for the DLF Steering Committee. 17 May 2005: http://www.diglib.org/architectures/serviceframe/dlfserviceframe1.htm) ; IMF (IMS Global Learning Consortium): zie http://www.imsglobal.org/ (o.a. IMS Global Learning Consortium (2003). IMS Digital Repositories v1.0 Final specification. IMS, 30 Jan. 2003: http://www.imsglobal.org/digitalrepositories/).

toepassingsbereik binnen archiefsystemen, de efficiëntie en helderheid van het OAIS-model en de focus ervan op langetermijnbewaring verantwoorden echter onze keuze voor OAIS.[2] Voor de toelichting van het OAIS-model baseren we ons grotendeels op het CCSDS-rapport van 2002.[3]

2.2 Open Archief Informatie Systeem

Het OAIS-referentiemodel bestaat uit drie delen. Het eerste deel beschrijft de verantwoordelijkheden van een open archiefsysteem. Een gemeenschappelijke, abstracte terminologie wordt opgebouwd, die gebruikt kan worden binnen archiefomgevingen om alle facetten van de gebruikte systemen, procedures, informatiedragers en uitwisselingspakketten te beschrijven. Het tweede deel beschrijft een functioneel model van een OAIS-archief met alle werkprocessen die nodig zijn voor de langetermijnbewaring van data. Het derde en laatste deel betreft een informatiemodel voor de beschrijving van de opgeslagen, digitale data.

2.2.1 Verantwoordelijkheden OAIS-archief en terminologie.

Een mogelijke definitie van een OAIS-archief luidt: “een organisatie van mensen en machines die de verantwoordelijkheid op zich hebben genomen om informatie te archiveren en beschikbaar te stellen voor een doelpubliek, verder aangeduid met Designated Community”. Deze definitie beklemtoont de twee primaire taken van een archief:

1. het moet informatie voor lange termijn kunnen archiveren en
2. het moet deze informatie ter beschikking kunnen stellen van een speciale doelgroep, de Designated Community.

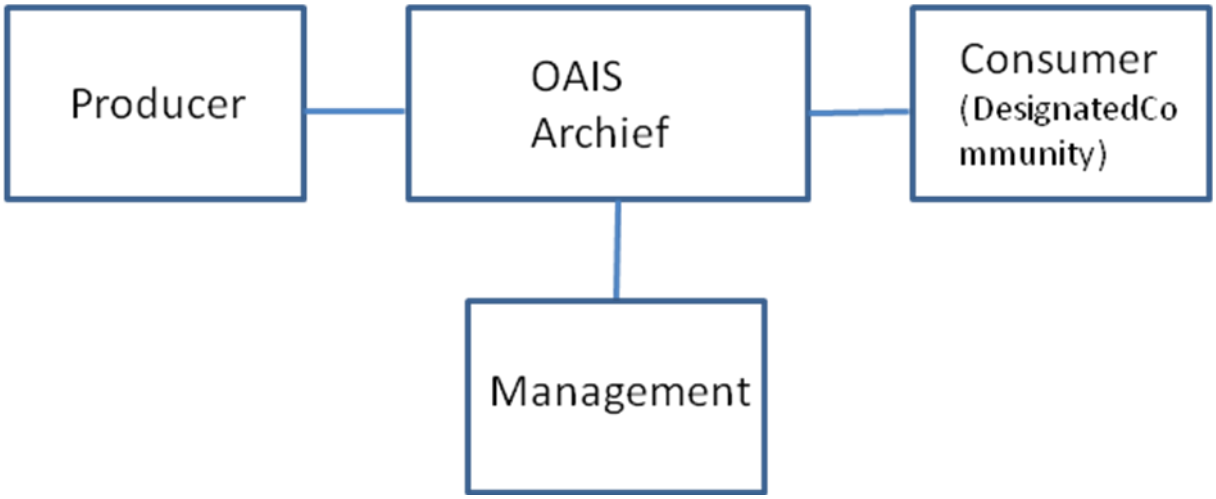
---

2 Ook J. Allinson in ‘OAIS as a reference model for repositories. An evaluation’, UKOLK, University of Bath, 21 november 2006, 17p. licht heel helder de bruikbaarheid van OAIS voor preserveringsprojecten en digitale repositories toe (http://www.ukoln.ac.uk/repositories/publications/oais-evaluation-200607/Drs-OAIS-evaluation-0.5.pdf). Ze refereert daarbij eveneens naar alternatieve referentiemodellen en suggereert dat ‘[e]valuating these would be a useful follow-on exercice’ (p. 14). Ten slotte wijst ze ook op enkele ‘Reference Models projects’ die door het JISC (Joint Information Systems Committee) zijn ingericht en die interessant kunnen zijn voor ‘the development of reference models for repositories’ (p. 14). Andere referenties die onze keuze voor OAIS ondersteunen, zijn: L. Brindley, ‘Taking the British Library forward in the twenty-first century’, in D-Lib Magazine, 6 (2000) 11. Online: http://www.dlib.org/dlib/november00/brindley/11brindley.html [aanhaling van OAIS-implementatie in de British Library]; K. Thibodeau, ‘Building the archives of the future. Advances in preserving electronic records at the National Archives and Records Administration’, in D-Lib Magazine, 7 (2001) 2. Online: http://www.dlib.org/dlib/february01/thibodeau/02thibodeau.html [verantwoording van het gebruik van OAIS in ERA, de Electronic Records Challenge, een initiatief van de National Archives and Records Administration (http://www.archives.gov/era/), waarin de mogelijkheden voor langetermijnbewaring van digitale records onderzocht worden]; N. Beagrie, ‘The continuing access and digital preservation strategy for UK Joint Information Systems Committee (JISC)’, in D-Lib Magazine, 10 (2004) 7-8. Online: http://www.dlib.org/dlib/july04/beagrie/07beagrie.html [toelichting bij de implementatie van OAIS door JISC.]. Voor andere voorbeelden van projecten, cf. §6.

3 CCSDS (2002). Reference Model for an Open Archival Information System (OAIS), CCSDS. Online: http://public.ccsds.org/publications/archive/650x0b1.pdf.

De Designated Community staat centraal in OAIS. Het is de doelgroep voor wie informatie gearchiveerd wordt. Een OAIS-archief moet ervoor zorgen dat deze groep gebruikers op elk moment de opgeslagen informatie kan raadplegen en begrijpen zonder een beroep te moeten doen op externe expertise. Alleen door het bereik van de Designated Community te bepalen, kan een OAIS garanderen dat de opgeslagen informatie voor lange termijn beschikbaar en begrijpelijk kan blijven.

De Designated Community kan het grote publiek zijn, maar dat is zeker geen vereiste. Het begrijpelijk houden van opgeslagen informatie voor een groot publiek, zonder dat men de dataexperts kan raadplegen, kan namelijk nodeloos een onoverkomelijk grote opdracht voor archieven vormen. We nemen als voorbeeld een OAIS-archief dat wetenschappelijke publicaties over een bepaald vakgebied bevat. De Designated Community bestaat dan mogelijk uit alle experten binnen deze discipline voor wie het archief een basis voor verder onderzoek vormt. Alle tabellen, meetresultaten, enz. moeten interpreteerbaar blijven voor deze Designated Community, zonder dat deze een beroep moet doen op Producers (cf. infra). Daarnaast kan het OAIS-archief deze informatie ook beschikbaar maken voor het grote publiek, zonder de strengere verplichting tot begrijpelijkheid.



Informatie wordt via een Ingest proces door een Producer aan een OAIS-archief geleverd. De interactie tussen een Producer en een OAIS-archief is vaak geformaliseerd in de vorm van een Submission Agreement, die de specifieke details van aanlevering bevat: welke datatypes worden aanvaard, welke metadatavelden moeten voorzien zijn, welke protocollen en logistiek moeten gebruikt worden, enz.

De beleidslijnen voor een OAIS-archief zijn bepaald door het Management. Deze groep is niet verantwoordelijk voor de dagdagelijkse werking van het archief. Management zet het beleid uit voor de te volgen archiefstandaarden, de strategische planning, de scope en draagt zorg en verantwoordelijkheid voor de beveiligde langetermijnbewaring van de aangeleverde informatie. Deze groep kan door middel van certificatie een trusted OAIS-archief uitbouwen.

Een OAIS-archief stelt een Consumergroep de opgeslagen informatie ter beschikking, waarbij de Designated Community een speciaal type ‘Consumer’ is. Het archief heeft de

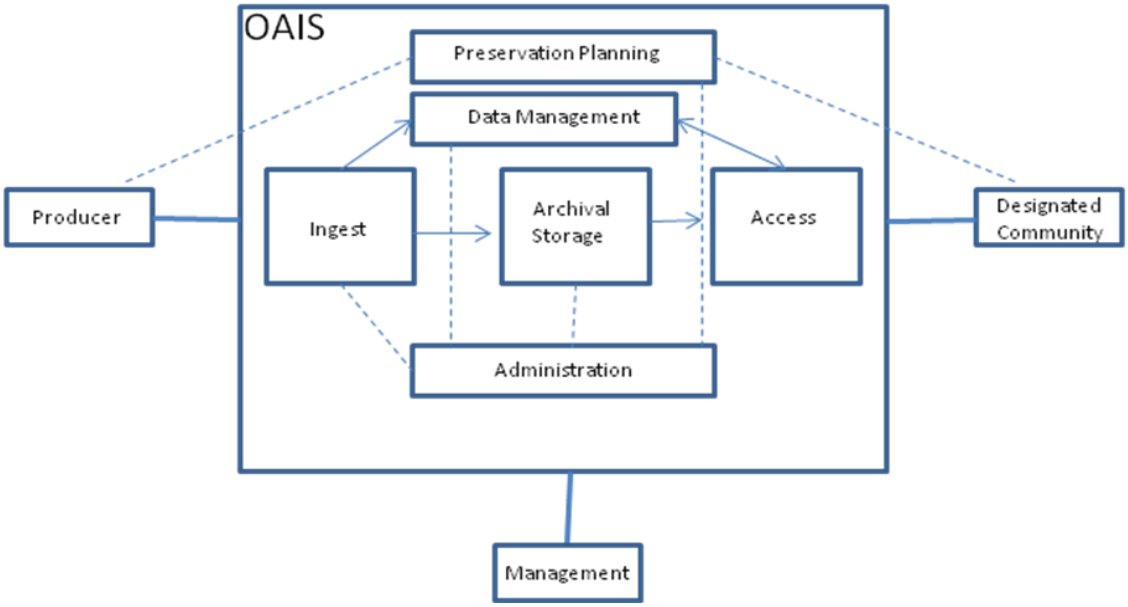
opdracht de opgeslagen informatie op zodanige wijze te archiveren dat de Designated Community onafhankelijk van de Producer de opgeslagen informatie kan interpreteren. Andere Consumers kunnen het grote publiek zijn, interne werkprocessen of externe OAIS-archieven die eventueel in een federatie samenwerken.

Om te voldoen aan het OAIS-referentiemodel moet een archief:

- onderhandelen met de Producers om alle nodige informatie te verkrijgen voor de archivering van hun data,
- genoeg rechten krijgen op de gearchiveerde informatie om de langetermijnbewaring te garanderen,
- de scope definiëren van de Designated Community,
- garanderen dat de opgeslagen informatie begrijpelijk blijft voor de Designated Community zonder dat deze een beroep moet doen op verdere assistentie van de Producers,
- gedocumenteerde procedures en richtlijnen volgen die garanderen dat de opgeslagen informatie gevrijwaard is van alle mogelijke risico's die beschikbaarheid of begrijpelijkheid van de opgeslagen informatie onmogelijk kunnen maken,
- op elk moment toegang kunnen verlenen tot de authentieke kopieën van de gearchiveerde informatie in originele vorm, of via een traceerbaar pad de originele vorm aanwijzen,
- de gearchiveerde informatie beschikbaar stellen aan de Designated Community.

2.2.2 OAIS Functioneel Model

In aanvulling op de bovenstaande gebruikersgroepen definieert OAIS een functioneel model voor de werking van een archief. In de bovenstaande tekst werd al het Ingestproces aangehaald waarmee Producers informatie aan een archief leveren. De Ingestmodule is de externe toegang tot het OAIS-archief die Producers zien. Specifieke functies zorgen voor de aanlevering van informatie, bevestiging van ontvangst, validatie van alle datacomponenten, transformatie van de informatie in een vorm die geschikt is voor de opslag en het beheer binnen het systeem, extractie of aanmaak van descriptieve metadata om de zoekinterfaces van het OAIS-archief te ondersteunen en het transport van de aangeleverde data naar de uiteindelijke archiefomgeving.



Een tweede component is de Archival Storage die de langetermijnbewaring van de gedigitaliseerde informatie garandeert. Deze component zorgt ervoor dat de geleverde data in geschikte vorm op online-, nearline- of offlinesystemen worden opgeslagen. Individuele bitstreams worden zodanig opgeslagen dat alle bits continu beschikbaar blijven zonder risico op bit-rot. De weergave van de bits in een presentatievorm moet ook gegarandeerd worden. Om aan beide eisen te voldoen, zullen regelmatig datamigraties naar nieuwe opslagmedia moeten plaatsvinden en zullen 'error checking-procedures' en 'disaster recovery' voorzien moeten zijn. Dataformaten zullen wellicht geconverteerd moeten worden naar nieuwe formaten indien technische wijzingen plaatsvinden die ervoor zorgen dat de bitstreams niet meer afgespeeld kunnen worden met standaardtools. Door software-emulatie kunnen bitstreams behouden blijven indien dataconversie onmogelijk is wegens het risico op informatieverlies of wegens het gebrek aan voldoende 'Representation Information' (cf. infra).

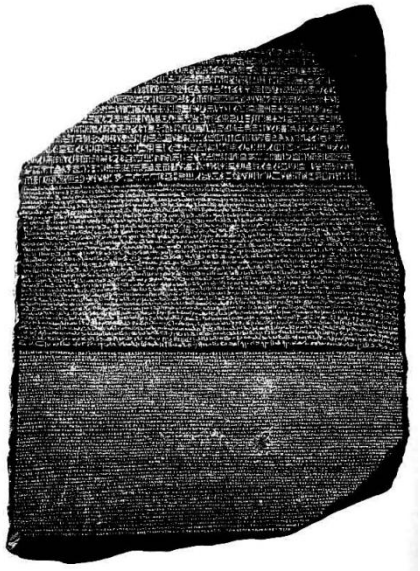
Een derde component, 'Data Management', verzorgt de catalogusomgeving waarbinnen de gearchiveerde informatie geïdentificeerd en beschreven wordt. De catalogus bevat naast de administratieve metadata, die de interne werking van het OAIS-archief ondersteunen, beschrijvingen van alle versies van opgeslagen files en systemen, de geschiedenis van de datamigraties en formaatconversies. 'Data Management' verzorgt rapporteringen en service requests van componenten binnen het OAIS-archief.

Een vierde component is 'Preservation Planning', die op lange termijn instaat voor het onderhoud van de gearchiveerde data. De component of dienst controleert de staat van de opgeslagen informatie: is deze nog steeds leesbaar met behulp van de huidige technologieën en is de informatie nog begrijpelijk voor de Designated Community? De dienst schat de impact van veranderende technologieën op de gearchiveerde informatie in en stelt een planning op om het OAIS-archief aan zijn verplichtingen tegenover de Designated Community te laten voldoen. Er worden strategieën ontworpen voor eventuele migratie, conversie of emulatie van de informatie en er wordt gezorgd voor de implementatie van deze strategieën in het OAIS-archief.

De vijfde component, 'Access', verzorgt de toegang tot het OAIS-archief: in enge zin voor de Designated Communty, in ruime zin voor de gebruikersgroep Consumers. De component stuurt zoekvragen van de Consumer door naar het Data Management en presenteert de opgeleverde metadata, eventueel geconverteerd in een vereenvoudigde vorm. Access is ook verantwoordelijk voor de authenticatie en autorisatie van eindgebruikers en toegang tot gearchiveerd materiaal. Gearchiveerde data uit de Archival Storage worden (eventueel na een interne conversie) in een presenteerbare vorm doorgegeven aan de eindgebruikers. Deze component heeft als primaire taak de gearchiveerde data toegankelijk te maken voor de Designated Community.

De laatste component, Administration, staat in voor de dagdagelijkse werking van het OAIS-archief. Administration verzorgt niet alleen de contacten met de Producers en Consumers, maar is ook verantwoordelijk voor de archief- en toegangssystemen. Operaties zoals monitoring, system performance en updates worden door deze component verzorgd. De Administration staat in direct contact met alle andere OAIS-componenten.

2.2.3 OAIS-Informatiemodel



Figuur 1 De Steen van Rosetta

Inleiding

Om langetermijnbewaring van informatie mogelijk te maken, is een duidelijke definitie van informatie in het kader van OAIS noodzakelijk.

Personen of systemen hebben een Knowledge Base die hen toelaat om een set aangeleverde informatie te begrijpen. Zo kan iemand met kennis van het hiëroglyfenschrift en de oud-Egyptische taal oude Egyptische teksten lezen en interpreteren.

De definitie van Information luidt: 'elk type kennis dat uitgewisseld kan worden in de vorm van data'. Data vormen een representatie van de informatie. Hiernaast is een oud-Egyptische tekst (informatie) als hiëroglyfen (data) gerepresenteerd op de Rosettasteen. De combinatie van de hiëroglyfen en de kennis van de oude Egyptische taal en haar schrift vormt betekenisvolle, begrijpelijke informatie. Zonder kennis van deze taal of dit schrift kunnen de data niet geïnterpreteerd worden. De data moeten begeleid worden door een beschrijving van het Egyptische schrift en door de wijze waarop men de tekens moet omzetten naar een taal die wel tot de Knowledge Base behoort, bijv. het oud-Grieks. Deze begeleidende data heten Representation Information. Iemand met een Knowledge Base die het oud-Grieks bevat, kan nu de Rosettasteen begrijpen. Dit is wat in werkelijkheid gebeurde met de Rosettasteen aan het begin van de 19e eeuw toen Jean-François Champollion het oud-Egyptische schrift kon ontcijferen met behulp van zijn kennis van het oud-Grieks.

Bovenstaande definities van Information, Data en Representation Information zijn ook toepasbaar op digitale informatie. Een voorbeeld: een file van 50MB (Data) wordt geleverd aan een OAIS-archief. Deze Data zijn een bitstroom die zonder Representation Information niet interpreteerbaar is. Begeleidende data moeten deze bitstroom beschrijven als een TIFF 6.0 bestand, dat een afbeelding bevat in de Adobe RGB 1998 kleurenruimte en dat een scan betreft met de afbeelding van de Rosettasteen met Griekse en Egyptische teksten. Om de bitstroom voor lange termijn toegankelijk te maken, moet een OAIS-archief in papieren of digitale vorm informatie opslaan over de TIFF 6.0 standaard, Adobe RGB 1998 kleurenruimte en in extreme vorm moet ze ook kennis over de taal waarin het object is opgesteld, meegeven. Deze laatste eis kan van toepassing zijn in bepaalde vakgebieden waarin het gebruikte jargon aan verandering onderhevig is. In zulke gevallen is de archivering van een woordenlijst noodzakelijk om de opgeslagen informatie binnen een wetenschappelijk discours te bewaren.

De recursieve aard van dergelijke pakketten is een bijkomende complicatie. Representation Information zoals de TIFF 6.0 standaard, ZIP, Adobe RGB 1998, Grieks, is ook een vorm van Data die gearchiveerd en beschreven moet worden met eigen Representation Information. Dat leidt tot een netwerk van beschrijvingen van beschrijvingen. Een OAIS-archief moet de Knowledge Base van een Designated Community kennen om archivering van Representation Information minimaal te houden. Bovendien kan de Knowledge Base van de Designated Community over lange tijd evolueren waardoor de nodige Representation Information moet aangepast worden.

In de praktijk kan men er niet van uitgaan dat de bewaring van Representation Information niet nodig zou zijn omdat er altijd sofware beschikbaar is om de dataobjecten te renderen. Dat is namelijk een illusie. De archivering van een werkende IT-infrastructuur (software & hardware), bijvoorbeeld door middel van emulatie, vormt een groter probleem dan de opslag van Representation Information in digitale of papieren vorm.[4]

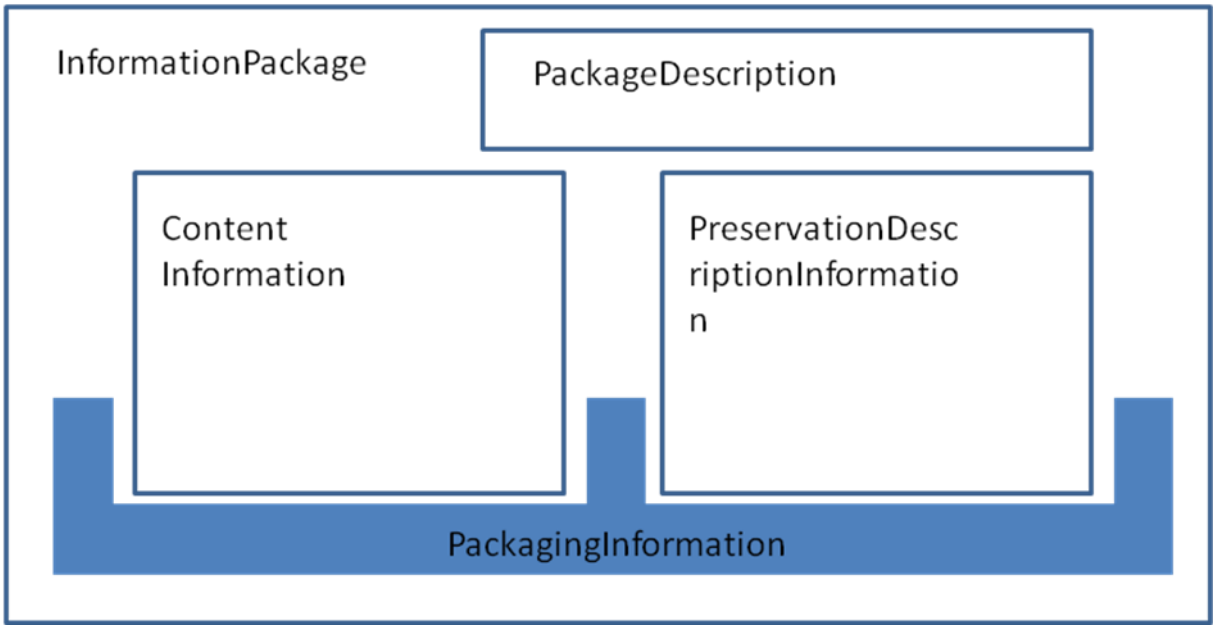
De bovenstaande beschrijvingsmethode van data tot op bitniveau lijkt tegengesteld aan objectgeoriënteerde werkwijzen waarbij deze implementatiedetails worden verborgen. Dit is echter een typische vereiste bij de archivering van digitale data. Digitale informatie is een conceptueel object dat altijd geïnterpreteerd moet worden binnen een specifieke IT-infrastructuur die aan veel technologische veranderingen onderhevig is. Bovendien wordt deze informatie ook altijd door mensen en organisaties geïnterpreteerd. In het laatste geval zullen ook de kennisdomeinen en organisatiestructuren zich over lange termijn wijzigen. Dat zorgt ervoor dat digitale informatie geen statisch gegeven is, maar voortdurend moet worden getoetst aan de gangbare praktijk.

4 Cf. de auteurs van het rapport CCSDS (2002), p. 2-4: ‘it is harder to preserve working software than to preserve information in digital or hardcopy forms’.

Informatiepakketten

Voor de uitwisseling van data tussen Producers, het Archive en de Consumers voorziet OAIS Information Packages. Deze pakketten zijn containers waarbinnen twee types informatie worden opgeslagen:

- Content Information, die de te archiveren informatie bevat,
- Preservation Description Information, die metadata bevat voor de langetermijnarchivering van Content Information.



Zoals in de inleiding beschreven, is Content Information zelf opgebouwd uit twee delen:

- Data, de representatie van informatie die gearchiveerd moet worden (bijv. een beeldbestand),
- Representation Information, de informatie die Data omvormt tot interpreteerbare concepten (bijv. de TIFF standaard die beschrijft hoe een reeks bytes omgevormd kan worden tot een beeld).

Representation Information bevat twee types data:

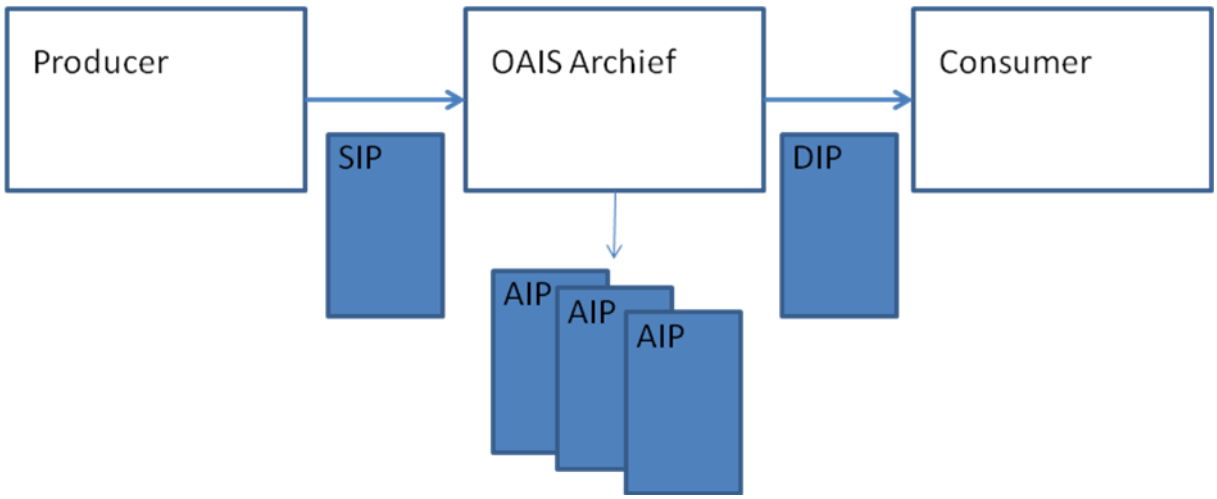
- Structure Information beschrijft hoe bytes omgezet kunnen worden in interpreteerbare concepten zoals letters, pixels, geluid, tabellen, enz.
- Semantic Information is een uitbreiding van Structure Information. Zelden volstaat informatie over een gebruikte standaard (bijv. TIFF) om de data te interpreteren. Semantic Information geeft een beschrijving van de informatie zelf: bijvoorbeeld de taal van een document, de beschrijving van een beeld, de relaties tussen de dataobjecten.

Voor de beschrijving van Content Information wordt een Preservation Description Information (PDI) pakket gevormd dat de volgende data bevat:

- Provenance beschrijft de ontstaansgeschiedenis van de Content Information: wie had de data in zijn bezit, via welke processen zijn de data tot de huidige vorm gekomen en welke versies zijn er beschikbaar. Voorbeelden van Provenance-metadata zijn:
    - Beschrijvingen van de scanapparatuur
    - Metadata over het scanproces en de gebruikte software en processen
    - Pointers naar het analoge origineel
    - Versiegeschiedenis van alle files
    - Copyright statements, beschrijving van licentiehouders
- Context beschrijft welke relaties de Content Informatie heeft met informatie die niet in het Informatiepakket zit. Voorbeelden van Context-metadata zijn:
    - Gerelateerde datasets
    - Pointers naar documenten in de originele omgeving op het moment van publicatie
    - Helpfiles
    - Taal
- Reference beschrijft de externe en interne identifiers waarmee Content Information op een unieke wijze geïdentificeerd kan worden. Voorbeelden van Reference zijn:
    - Object identifiers
    - Bibliografische beschrijvingen
    - ISBN's, ISSN's, DOI's, Handlers
    - Versienummers
    - Namen, titels
- Fixity bevat checksums en andere beveiligingen die testen of Content Information op een ongedocumenteerde wijze werd aangepast. Voorbeelden van Fixity zijn:
    - Checksums
    - Digital signatures
    - Certificaten
    - Encryptie
    - CRC's

Package Information is de metadata die alle onderdelen van de Content Information en Preservation Description Information op logische of fysieke wijze aan elkaar binden. Wanneer bijvoorbeeld de Content Information en Preservation Description Information in een ZIP-file worden aangeleverd, dan is de Package Information de Manifest file die de filenamen en beschrijvingen bevat van alle files in het ZIP-pakket.

Package Description Information bevat de metadata om bepaalde Content Information Packages terug te vinden in grote collecties, bijvoorbeeld via een Dublin Core-record.



De bovenstaande Information Packages worden uitgewisseld tussen de Producer, het Archive en de Consumer. In realiteit bevat niet elk pakket voldoende archiefmetadata om aan de OAIS-standaarden te voldoen. Zo bevatten door Producers aangeleverde pakketten doorgaans onvoldoende Preservation Description Information en de opbouw van de pakketten is wellicht niet onmiddellijk geschikt voor opslag in het langetermijnarchief. Information Packages die aan Consumers worden geleverd zullen dan weer minder of een ander type data bevatten dan in het archief aanwezig is.

OAIS onderscheidt drie types Information Packages:

- Het Submission Information Package (SIP) is het Information Package dat door de Producer aan het OAIS-archief wordt geleverd,
- Een of meerdere SIP's vormen samen, na de nodige transformaties en metadataverrijking, een Archival Information Package (AIP), dat in het archief wordt opgeslagen,
- Op antwoord van vragen door Consumers zal het Archief een AIP of delen daarvan, getransformeerd of niet, vrijgeven als een Dissemination Information Package (DIP).

2.2.4 Metadatamodellen

Op basis van het OAIS-referentiemodel zijn er diverse metadatatypes nodig om een volledige archiefbeschrijving te geven van informatie. In de onderstaande tabel wordt een overzicht gegeven van bestaande metadatamodellen, waarvan de belangrijkste verder in dit rapport beschreven worden:

Descriptieve Metadata	MARC, MODS, Dublin Core, P-META, ISAD, VRA/Core, SMEF-DM, EAD, XMP, IPTC, FRBR, CDWA, Object Id
Vocabulaires/ Thesauri	LCSH, Getty, RAMEAU, MESH, Dewey, GILS, AAT, FIAT/IFTA, SKOS, IUPAC, BFO, CIDOC, ULAN, ISAAR(CPF), Thesaurus architecture et patrimoine, TGN

Package Information	METS, MPEG-21/DIDL, IMS-LOM, ORE, WARC
Technische Metadata	MPEG-7, EXIF, AudioMD, VideoMD, TextMD
Referentie Metadata/Identifiers	URI, URN, URL, Handle, DOI, Purl, INDO, NBN, OpenURL, Ark, ISSN, ISBN, …
Provenance	PREMIS, VERS, RKMS, ISO-23081
Fixity	XMLSignatures, MD5, SHA, …
Data formaten	MPEG, H264, DivX, DIRAC, MPJEG, Theora, MP3, AAC, FLAC, Ogg, TTA, WMA, JPEG, JPEG2000, GIF, PNG, TIFF, WAV, AIFF, XMF, Matroska, MXF, MP4, 3GP, ASF, MOV, AVI, FLV, RealMedia

3 Dataformaten

3.1 Inleiding

Het opslaan en versturen van multimedia (beeld, geluid, video) in zijn ruwe vorm is vandaag de dag meestal niet meer wenselijk. De benodigde hoeveelheid opslagruimte en bandbreedte is immers amper te becijferen. De technologische revolutie die hiervoor een antwoord bood is broncodering. Broncodering heeft immers als doel het efficiënt representeren van informatie, zodat er op een optimale wijze gebruik kan worden gemaakt van schaarse middelen zoals opslagcapaciteit en transportcapaciteit. Bij broncodering wordt altijd getracht maximaal gebruik te maken van eventuele redundantie die aanwezig is in de te transporteren of te bewaren informatie. Daarnaast wordt bij sommige vormen van broncodering ook getracht gebruik te maken van beperkingen bij de ontvanger van de informatie. In de informatie aanwezige subtiliteiten, die een ontvanger ontgaan, moeten immers niet worden opgeslagen of getransporteerd. De gepaarde algoritmes die ontwikkelt worden om de ruwe informatie te coderen/comprimeren en daarna te decomprimeren/decoderen, worden een codec genoemd.

De meest gebruikte videocompressiestandaard vandaag de dag is nog steeds MPEG-2. Met de opkomst van performante eindgebruikerstoestellen voor het opnemen en afspelen van videomateriaal, de alomtegenwoordigheid van breedbandverbindingen en de intrede van Hoge-Definitie TV en DVD (HD-DVD & Blu-ray) zal H.264/MPEG-4 AVC de fakkel op korte termijn zeker overnemen. Digitale cinema en professionele postproductie daarentegen behouden hun vertrouwen in Motion JPEG 2000. De “open source” wereld stelt daar tegenover de patentvrije Theora dat probeert te wedijveren met H.264/MPEG-4 AVC en Motion JPEG 2000.

De populairste huidige audiocompressietechniek is zonder twijfel MP3. Desalniettemin zal deze op korte termijn zeker vervangen worden door AAC omdat deze een veel betere kwaliteit garandeert bij eenzelfde bitrate. Voor langdurige preservatie van audio zal er evenwel veelal gekozen worden voor verliesloze compressie (vb. FLAC) of zelfs ongecomprimeerde opslag (zuivere PCM samples) zoals nu reeds soms wordt toegepast op de nieuwe Blu-ray schijfjes omdat de hoeveelheid informatie die moet opgeslagen worden voor audio een paar grootteordes verschilt van video. Waar de opgeslagen hoeveelheid informatie bij video exponentieel stijgt naarmate grotere resoluties moeten bewaard worden, is deze hoeveelheid informatie bij audio enkel afhankelijk van het aantal kanalen (stereo, 5.1, …) dat men wenst te bewaren. Digitale cinema is hier dan ook de slokop met hun Dolby TrueHD standaard die tot 14 kanalen (13.1) aan kan. De huidige Blu-ray en HD-DVD standaarden ondersteunen tot op heden slechts 8 kanalen (7.1).

JPEG is veruit de meest algemeen verspreide compressiestandaard voor afbeeldingen. Het werd specifiek ontworpen om digitale beelden te comprimeren en is op dit moment dé standaard voor afbeeldingen op het internet en bij digitale camera's. Binnen de wereld van de digitale cinema is zijn superieure opvolger JPEG2000 echter "de facto" standaard. Als verliesloze compressie een vereiste is, dan zijn PNG en TIFF de toonaangevende standaarden. Het is waarschijnlijk dat PNG in de internetwereld verder ingang zal vinden ten koste van JPEG. Daar waar kwaliteit echter primordiaal is, vb. bij allerhande archiveringstoepassingen, zal vooral TIFF gebruikt worden onder meer ook omdat het door alle platformen het beste ondersteund wordt.

Het opkomende containerformaat om A/V-materiaal en bijhorende data en metadata uit te wisselen in de wereld van de digitale cinema en omroepen is ongetwijfeld MXF. Het is ontworpen om content tijdens het productieproces probleemloos te kunnen uitwisselen. In de internetwereld zal AVI en WMV, vanwege de Microsoft dominantie op OS-niveau, nog een tijd als A/V-containerformaat de dienst uit maken. De superieure codec H.264 die op het internet en in de mobiele wereld furore maakt, heeft echter ook een eigen containerformaat MP4 waardoor zijn belangrijkheid als A/V-container in de toekomst alleen maar zal toenemen. Om A/V-materiaal te archiveren worden vooralsnog vooral AVI en in de toekomst zeker ook MXF als containerformaten gebruikt daar ze beiden ook ruwe data (dus geen compressie) kunnen herbergen.

3.2 Compressieformaten

3.2.1 MPEG

MPEG is een verzameling standaarden voor de compressie van beeld- en geluidsbestanden.

Situering

In 1988 werd de werkgroep Moving Pictures Experts Group opgericht als een samenwerkingsverband tussen academici en mensen uit de bedrijfswereld. Sindsdien houdt deze werkgroep zich bezig met de ontwikkeling van standaarden voor de codering van audio en video.

De MPEG-codecs zijn een voorbeeld van datacompressie met gegevensverlies ("lossy compression"). Dat is een methode waarbij decompressie van een gecomprimeerd bestand resulteert in een bestand dat verschilt van het origineel. Er zal verlies van informatie optreden, met kwaliteitsverlies tot gevolg.

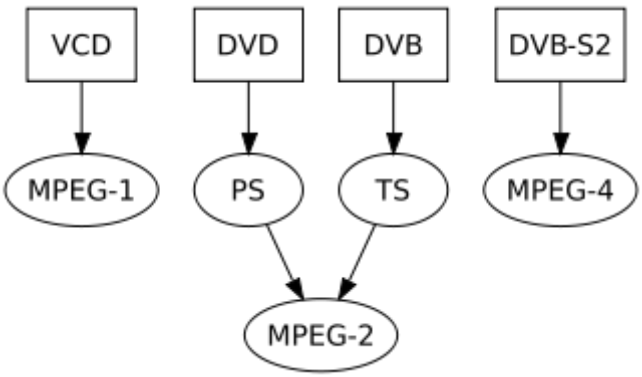
De lossy MPEG-compressiealgoritmes maken gebruik van een transformatietechniek waarbij beelden en geluiden in kleine segmenten verdeeld worden, vervolgens getransformeerd worden naar het frequentiedomein om ten slotte gekwantiseerd te worden.

Meestal wordt er een extra stap toegevoegd, waarbij een bepaalde afbeelding voorspeld wordt op basis van de vorige geconstrueerde afbeeldingen. Enkel de verschillen met de vorige afbeeldingen en de extra informatie om de voorspelling te kunnen vormen, worden bijgehouden. Deze voorspellingsmethode wordt ook wel gebruikt in andere compressieformaten.

MPEG standaardiseert enkel het bitstreamformaat en de decoder. Een bitstreamformaat is de vorm waarin de data zich bevinden, namelijk als een reeks van bits die gebruikt worden in een toepassing voor digitale communicatie of opslag. De encoder is niet gestandaardiseerd, maar er zijn implementaties beschikbaar voor hen die gevalideerde bitstreams produceren. Zo kan bijvoorbeeld een willekeurige MPEG-decoder om het even welk MPEG-materiaal van hetzelfde type decoderen, zonder rekening te moeten houden met de encoder.

Formaten

MPEG heeft een aantal compressieformaten gestandaardiseerd. Elk formaat heeft daarbij een onbekend aantal codecs. Aangezien de bitstream en de decoder gestandaardiseerd zijn, kan een decoder van een bepaald formaat iedere MPEG-bitstream van dat formaat decoderen. Het is voor de decoder dus niet van belang met welke codec een bepaald bestand geëncodeerd werd.



MPEG-1

MPEG-1 (1991) is de eerste compressiestandaard voor video en audio die door de Moving Picture Experts Group ontwikkeld werd. Deze werd later gebruikt als standaard voor het video-cd formaat (VCD). Dit formaat beschrijft ook het populaire Layer 3 (MP3) audiocompressieformaat. De videocodec is enkel van toepassing voor niet-geïnterlinieerde

beelden. Het formaat beschrijft verder ook nog synchronisatie en multiplexing van video en audio, procedures om de conformiteit te testen en referentiesoftware.

MPEG-2

Het MPEG-2-formaat is voornamelijk ontwikkeld voor het transport van digitale kwalitatieve video en audio voor televisie-uitzendingen. Het wordt gebruikt voor digitale televisie via conventionele antennes (dus niet via satelliet) (ATSC, DVB en ISDB), uitzending ('broadcasting') via satelliet (DirecTV) en digitale kabeltelevisie. Mits een kleine aanpassing is het ook toepasbaar op dvd-videodiscs (DVD).

Volgens het MPEG-2-formaat kunnen onbewerkte frames in drie soorten frames gecomprimeerd worden:

- intragecodeerde frames (I-frames)
- voorspellend gecodeerde frames (Predictive coded, ofwel P-frames)
- bidirectionele P-frames (B-frames)

Meestal wisselen I-, P- en B-frames af. Een mogelijke volgorde is: IBBPBBPBBPBB(I). De frames samen vormen een GOP (Group Of Pictures). Hoe de variatie moet optreden, is niet specifiek vastgelegd: de standaard is hier vrij flexibel in.

MPEG-3

Oorspronkelijk werd MPEG-3 ontworpen voor High-definition television (HDTV). Het bleek echter dat kleine aanpassingen aan MPEG-2 hetzelfde resultaat kunnen opleveren. De verdere ontwikkeling van MPEG-3 werd daarom beëindigd.

MPEG-4

MPEG-4 is een uitbreiding van MPEG-1 als ondersteuning voor video/audio-“objecten”, 3D-inhoud, lage bitrate-encoding en Digital Rights Management (DRM). Als bestandsformaat koos de Moving Pictures Experts Group voor QuickTime, die door Apple ontwikkeld is. Microsoft zag liever zijn bestandsformaat in een aparte ISO-standaard opgaan en kwam daarom uit met een eigen versie van MPEG-4. Op deze gesloten en incompatibele Microsoft-variant van MPEG-4 is overigens het populaire DivX gebaseerd.

Softwareleveranciers kunnen, conform de ISO-standaard, hun eigen codecs en bijhorende encoders ontwikkelen als dit een meerwaarde kan betekenen voor de door hen aangeboden producten. Enkele voorbeelden hiervan zijn Apple en 3ivx. In samenwerking met het Joint Video Team (JVT) werd op basis van MPEG-4 Part 2 een geavanceerde videocodec ontwikkeld. Deze kreeg de naam H.264 (of MPEG-4 Part 10) maar dit is uiteraard enkel een referentie.

3.2.2 H.264/MPEG-4 AVC/MPEG-4 Part 10

H.264, MPEG-4 Part 10 of MPEG-4 AVC (Advanced Video Coding) is een digitale videocodec die een heel sterke compressie van videobeelden nastreeft.

Door de sterke ontwikkeling van het internet is ook het aantal internetdiensten sterk toegenomen. Omdat ook de opslagcapaciteit en processorsnelheden sterk zijn toegenomen, lijkt de nood aan een verbeterde codec niet zo groot. Toch brengt een sterkere compressie andere voordelen met zich mee, namelijk:

- een betere kwaliteit bij eenzelfde bandbreedte voor streaming video
- snellere downloadtijden van videobestanden
- nog betere kwaliteit op DVD's
- langere films op een DVD
- ...

Daarom besloten ITU-T VCEG en ISO MPEG hun krachten te bundelen en ze richtten het Joint Video Team (JVT) op, met het oog op de ontwikkeling van een nieuwe videocoderingsstandaard: H.264/MPEG-4 AVC.

De encoder voor H.264/MPEG-4 AVC videostreams is het open source pakket VideoLAN-x264. De volgende toepassingen maken al gebruik van de x264 encoder:

- Google Video
- MobileASL
- Speed Demos Archive
- TASvideo

Zoals eerder al beschreven, heeft elk MPEG-formaat zijn specifieke toepassing. Zo wordt MPEG-2 voornamelijk gebruikt door digitale televisiebroadcasters en zal H.264/MPEG-4 AVC vooral bij de gewone consumenten zijn plaats veroveren (draadloos netwerk). Naderhand zal H.264/MPEG-4 AVC echter ook op professioneel gebied MPEG-2 vervangen. Zo wordt bij iDTV & IPTV toepasssingen steeds meer H.264/MPEG-4 AVC geïmplementeerd in plaats van enkel MPEG-2. De grote winst aan bandbreedte bij het bekomen van eenzelfde kwaliteit is uiteraard een sterk argument.

Met de sterke opgang van H.264/MPEG-4 AVC, zal vooral het gebruik van MPEG-1 afnemen. VHS-kwaliteit volstaat voor de veeleisende consument niet langer en de uiterst efficiënte videocompressie van H.264/MPEG-4 AVC zorgt ervoor dat dit formaat de bovenhand zal krijgen.

Toepassingen

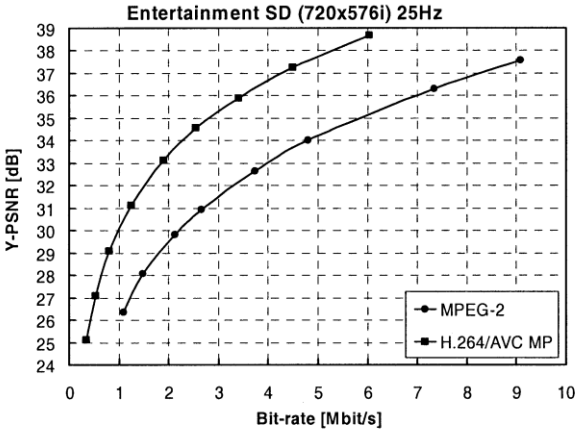
Om intercompatibiliteit te verzekeren, waarbij verschillende producten de standaard op dezelfde manier implementeren, werd de Internet Streaming Media Alliance (ISMA) opgericht als een samenwerkingsverband tussen Apple, Cisco, IBM, Kasenna, Philips, Sun Microsystems, AOL Time Warner, Dolby Laboratories, Sony en 27 andere bedrijven. De ISMA richt zich vooral op de definiëring van bruikbare profielen. Binnen H.264/MPEG-4 AVC werden aldus een 7-tal profielen gedefinieerd, elk met een specifieke groep van toepassingen voor ogen:

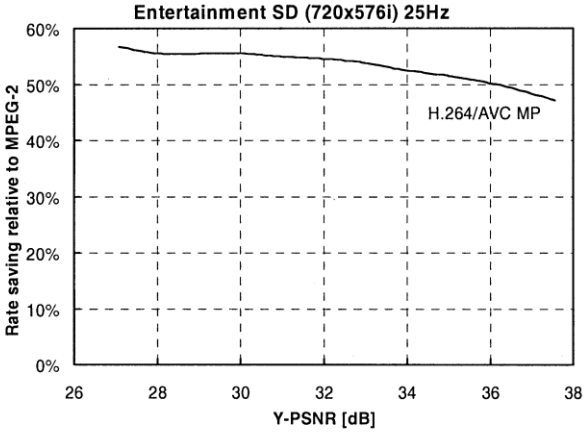
- **Baseline Profile (BP)**: voornamelijk bedoeld voor lage-kost toepassingen op toestellen met een beperking qua beschikbare resources, waarvan videoconferencing en mobiele toepassingen een typisch voorbeeld zijn.
- **Main Profile (MP)**: initiëel bedoeld als belangrijkste gebruikersprofiel voor broadcast applicaties en opslag. Dit profiel verliest echter aan belang sinds de uiteindelijke ontwikkeling van het “High Profile” dat eenzelfde soort applicaties ook ondersteunt.
- **Extended Proile (XP)**: bedoeld als streaming video profiel heeft dit een relatief hoge compressie mogelijkheid, bovenop een paar extra “robustness features” om te kunnen omgaan met dataverlies en het switchen van streaming servers.
- **High Profile (HiP)**: het belangrijkste profiel voor braodcast en opslag applicaties, vooral reeds in gebruik voor Hoge-Definitie televisie (HD). Dit profiel wordt immers reeds ondersteund voor de nieuwe generatie van DVD-formaten, met name HD-DVD en Blu-ray.
- **High 10 Profile (Hi10P)**: een toekomstgericht profiel dat verder bouwt op het “High Profile”, waarbij een precisie van 10 bits per sample van een gedecodeerd beeld mogelijk wordt.
- **High 4:2:2 Profile (Hi422P)**: een toekomstgericht profiel dat verder bouwt op het “High 10 Profile” gericht op professionele applicaties die interlaced video gebruikt, waarbij 4:2:2 subsampling mogelijk wordt bij een 10 bits per sample voor een gedecodeerd beeld.
- **High 4:4:4 Predictive Profile (Hi444PP):** een toekomstgericht profiel dat verder bouwt op het “High 4:2:2 Profile”, waarbij 4:4:4 subsampling mogelijk wordt bij een 14

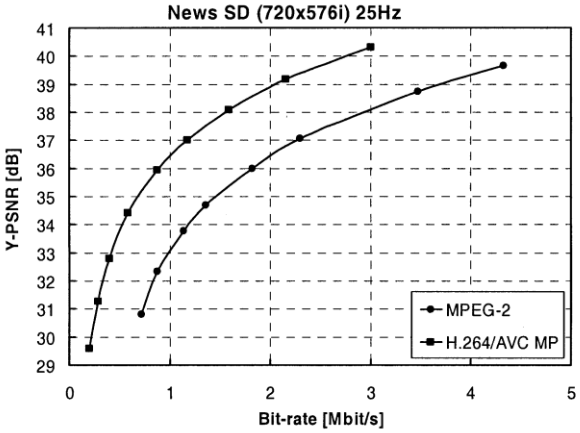
bits per sample voor een gedecodeerd beeld. Verder ondersteunt het ook “lossless region coding” en het apart coderen van elk beeld als 3 onderscheiden kleurvlakken.

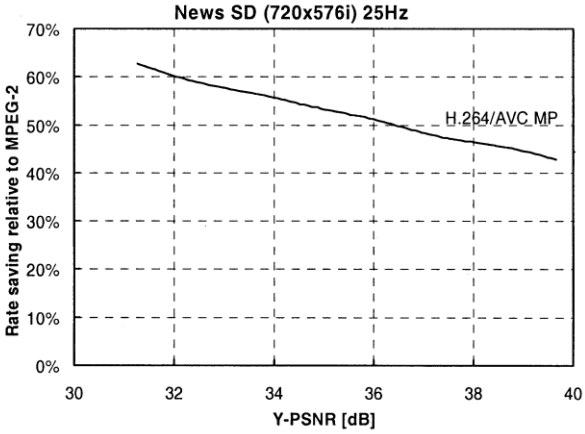
HD-TV

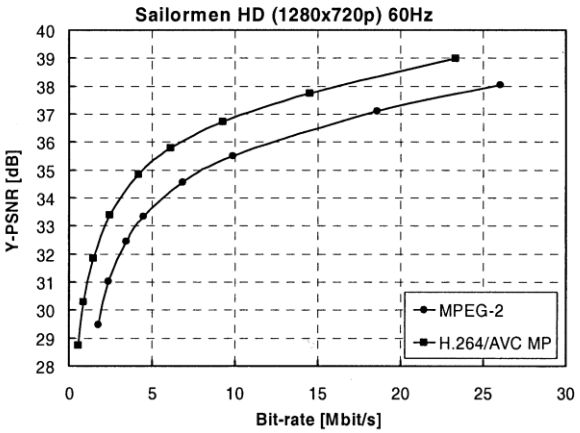
Binnen de broadcast wereld (waaronder ook IPTV en iDTV gerekend kan worden) zal de huidige MPEG-2 standaard voor Hoge-Definitie beelden op korte termijn vast en zeker vervangen worden door H.264/MPEG-4 AVC High Profile. De overschakeling van analoge televisie naar digitale televisie en de quasi gelijktijdige overgang van standaard televisie (SD-TV) naar Hoge-Definitie televisie (HD-TV) heeft de behoefte naar bandbreedte immers sterk doen stijgen. Met een huidige bandbreedteperformatiewinst van om en bij de 50% lijkt het lot van MPEG-2 bezegeld.

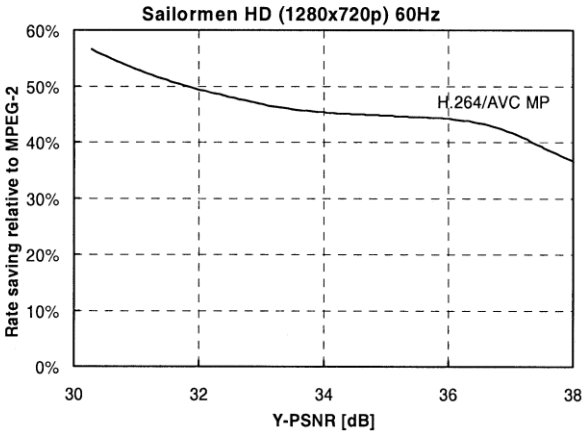












Players

Aangezien voorspeld wordt dat H.264/MPEG-4 AVC het meest gebruikte formaat voor video en audio zal zijn, wordt dit formaat al door de meeste nieuwe spelers ondersteund. Hieronder kunnen we onder andere QuickTime 6, RealPlayer 10 en de VLC media player opmerken. Microsoft weigert echter nog steeds om MP4-bestanden in Windows Media Player te ondersteunen.

Streaming

Het MPEG-formaat is uitermate geschikt voor streaming media. De Moving Picture Experts Group heeft daarvan altijd het nut ingezien en er rekening mee gehouden bij de ontwikkeling van hun formaten en implementatie van codecs. De industrie volgde deze redenering niet, waardoor MPEG op commercieel vlak nooit als streamingformaat doorgebroken is. Dit hield voornamelijk verband met het feit dat concurrerende formaten als RealMedia en Windows Media met hun eigen streamingserver op de markt kwamen. MPEG is immers een standaard die niet wordt gestuwd door een commercieel bedrijf en dus afhankelijk is van implementaties en ontwikkeling door derden.

Voor H.264/MPEG-4 AVC lijkt hier verandering in te komen. Een belangrijk voorbeeld hiervan is de Darwin Open Source Streamingserver van Apple, die streaming van H.264/MPEG-4 AVC ondersteunt. Alles zal echter afhangen van de ondersteuning door de players. Inmiddels is de QuickTimespeler, dankzij het succes van iPod en iTunes, vrijwel even bekend als de Windows Media-speler. Vooral in het mobiele segment en de Set-top-box (iDTV & IPTV) zijn de kansen voor MPEG-streaming aanzienlijk.

3.2.3 VC-1

VC-1 is een video codec specificatie die is gestandardiseerd door SMPTE (Society of Motion Picture and Television Engineers). De specificatie is reeds geïmplementeerd door Microsoft in Windows Media Video 9 (WMV 9). Bij de standardisatie van VC-1 waren 75 organisaties betrokken. SMPTE 421M beschrijft de bit stream syntax en wordt vergezeld door twee andere documenten (SMPTE RP227 en SMPTE RP228) die het transport en de conformiteiten van VC-1 behandelen.

SMPTE is een organisatie van film en video experten, met leden in 85 landen. De standaarden die SMPTE invoert, worden wereldwijd overgenomen door professsionals op het gebied van video, bewegend beeld en digitale cinema. De SMPTE stadaard voor VC-1, SMPTE 421M, was oorspronkelijk gebaseerd op de Windows Media Video 9 codec. Deze codec is in feite niets anders dan de Microsoft implementatie van de VC-1 standaard.

De redenen waarom Microsoft koos voor de standardisatie van VC-1 waren toegankelijkheid en interoperabiliteit. De standardisering stimuleert onafhankelijke ontwikkeling en verzekert dat deze implementaties interoperabel zijn. De standardisering van de bitstream syntax en het decoderingsproces zorgen voor de nodige bronnen en stabiliteit om te kunnen investeren in het creëren van decoders op chip. Door de isolatie van de video codec van de rest van het

video systeem kan deze in verschillende types hardware en in vele verschillende systemen worden ingebouwd. De standardisatie bevordert dus de adoptie van de technologie binnen de sector.

De VC-1 codec is ontwikkeld om een uitmuntende video kwaliteit te bereiken bij zeer lage tot zeer hoge bitsnelheden. De codec kan gemakkelijk overweg met 1920 x 1080 pixels beelden bij bitsnelheden van 6 tot 30 Mbps voor high-definition video. VC-1 kan ook hogere resoluties aan zoals b.v. 2048 x 1536 pixels beelden voor digitale cinema en heeft een maximum bistelheid van 135 Mbps. Een voorbeeld van zeer lage bitsnelheidsvideo zou een film zijn van 160 x 120 pixels bij een bitsnelheid van 10 Kbps voor modem applicaties.

VC-1 is een evolutie van de conventionele DCT-gebaseerde video codec ontwerpen zoals die zijn terug te vinden in H.261, H.263, MPEG-1, MPEG-2, MPEG-4 Part2 en AVC(MPEG-4 Part 10/H.264). Bijgevolg is VC-1 heel gelijkaardig aan H.264. Het bezit een heel gama aan geavanceerde coding tools, maar er zijn toch een aantal verschillen. Deze verschillen situeren zich voornamelijk in details van de gebruikte filters.

VC-1 presteert duidelijk beter dan MPEG-2 en MPEG-4 simple profile en is qua performantie vergelijkbaar met H.264. Alhoewel de compressie efficiëntie hetzelfde is bij VC-1 en H.264, is VC-1 iets minder complex vergeleken met het baseline profile van H.264. VC-1 wordt op dit moment vooral gebruikt binnen de PC-omgeving (WMV9), maar kan ook belangrijk worden in de netwerk-omgeving. Op dit moment wint VC-1 ook aan terrein in de filmindustrie waar reeds enkele films werden uitgebracht die werden geëncodeerd met VC-1 voor het afspelen in HD. VC-1 is ook een standaard voor compressie bij de HD-DVD en Blu-ray formaten.

3.2.4 DivX

DivX is een standaard om digitale videobestanden compact op te slaan door gebruik te maken van een compressiealgoritme dat geoptimaliseerd is voor videobeelden.

Geschiedenis

DivX begon aanvankelijk als een hack van de Microsoft MPEG-4 codec. Het Microsoft MPEG-4 codeerformaat week echter op belangrijke punten af van de officiële MPEG-4 implementatie Verder was deze codec ook beschermd zodat het niet mogelijk was om op hoge resolutie films te coderen. Afspelen kon daarentegen wel, zodat een verbeterde encoder, tegen een behoorlijke prijs, films met een hoge kwaliteit kon encoderen. Op die

manier trachtte Microsoft te laten uitschijnen dat misbruik van hun codec oninteressant was. De gekraakte codec kon wel hogekwaliteitfilms coderen en ging door het leven als DivX3. De ontwikkeling van DivX ging hand in hand met de ontwikkeling en de opkomst van webvideo.

Van een illegale naar een legale codec

DivX3 was dus illegaal, wat uiteraard problematisch was. Het bedrijf DivX Networks werd opgericht om een legale DivX codec te ontwikkelen die bovendien de beeldkwaliteit nog zou optimaliseren. Men begon met een referentie implementatie van de officiële MPEG-4 standaard, Momusys. Om de ontwikkeling te versnellen, werd het project “open source” gemaakt. Dit leverde de DivX4 codec op. Maar het doel om een betere beeldkwaliteit dan DivX3 te leveren, werd niet bereikt. Het bleek namelijk mogelijk de beeldkwaliteit van DivX3 sterk te verbeteren door o.a. gebruik te maken van variabele bitrates. Zodra DivX4 uitgebracht was, veranderde DivX Networks zijn strategie. De licentie van de broncode liet toe dat je de code kon overnemen en commercieel verder ontwikkelen. Daarom werd DivX een gesloten project. DivX Networks had ook de merknaam DivX verworven.

XviD, de Open source DivX

DivX als gesloten project wekte uiteraard wrevel op bij de “open source” programmeurs van DivX4. Zij startten een nieuw project, XviD, waarin men aan de toen bestaande DivX broncode zou verder werken. DivX Networks verbeterde de DivX4 codec en bracht DivX5 uit. Deze codec verbeterde de beeldkwaliteit flink. Ook de XviD-programmeurs wisten de DivX4 codec te verbeteren en evenaren inmiddels de beeldkwaliteit van DivX. De prestaties van de beide codecs zijn zo goed dat zij DivX3 ver achter zich laten. Steeds meer hardware hebben nu al een ingebouwde DivX en/of XviD codec.

3.2.5 DIRAC

Dirac is een “open source” videocodec die door de BBC ontwikkeld werd. De codec wil het multimedialandschap, dat door gesloten en vaak zelfs gepatendeerde videoformaten geplaagd wordt, een open standaard bieden. BBC overweegt om op termijn haar volledige beeldmateriaal in het Dirac-formaat aan te bieden om zo het gebruik van de Dirac codec te verzekeren. Dirac gebruikt door de BBC ingediende patenten, maar de BBC geeft iedereen die de Dirac codec wil gebruiken, gebruiksrechten op deze patenten. Momenteel is Dirac nog niet gebruiksklaar. De codec is genoemd naar de Britse natuurkundige Paul Dirac, en dit zonder specifieke reden volgens BBC.

3.2.6 MJPEG/Motion JPEG/Motion JPEG2000

MJPEG is een toepassing van de Joint Photographic Experts Group (JPEG). MJPEG is echter een methode waarbij ieder frame van een beeldsequentie naar het JPEG- of JPEG2000-formaat wordt omgezet om zo een 10:1 tot 20:1-compressie te bekomen. Daarbij wordt dus enkel intraframe codeertechnologie gebruikt die vergelijkbaar is met de I-frames van de videocodecs MPEG-1 en MPEG-2 die daarenboven ook nog aan interframe prdeictie doen. Doordat MJPEG geen interframe predictie ondersteunt, verliest het enige compressie mogelijkheden, maar maakt het video editeerproces gemakkelijker aangezien simpele editeeroperaties op alle beelden kunnen toegepast worden aangezien die allemaal I-frames zijn. Ook MPEG-2 kan op deze manier enkel met I-frames werken wat resulteert in eenzelfde editeergemak en compressiemogelijkheden verlies. Als men enkel intraframe codeertechnologie gebruikt, dan wordt de compressiemogelijkheid ook automatisch losgekoppeld van de hoeveelheid beweging in de scene, aangezien temporele predictie nu niet meer nodig is. Alhoewel de bitrate van MJPEG aanzienlijk beter is dan de bitrate van de volledig ongecomprimeerde video, is deze toch aanzienlijk slechter dan videocodecs die interframe bewegingscompensatie gebruiken, zoals MPEG-1 en MPEG-2. MJPEG wordt enkel en voornamelijk gebruikt in commerciële postproductie en digitale cinema.

3.2.7 Theora

Theora is een vrije “open source” codec voor videobestanden en wordt ontwikkeld door de Xiph.Org Foundation. Het is vrij van patenten en gebaseerd op de VP3-codec van On2 Technologies. Het is de bedoeling dat Theora kan wedijveren met bestaande codecs zoals MPEG-4 en WMV.

Techniek

Theora is een lossy compressiemethode en de gecomprimeerde video kan opgenomen worden in een geschikt containerformaat. Het wordt meestal gebruikt in het bestandsformaat Ogg, samen met de Ogg Vorbis audiocodec. Deze combinatie kan worden gebruikt voor de productie van een rechtenvrij multimediabestand. Andere codecs, zoals MPEG-4, zijn gepatenteerd en voor commercieel gebruik moet een licentie betaald worden. Theora maakt gebruik van chroma subsampling, op blokken gebaseerde compensatie voor beweging (block motion compensation, BMC) en een 8x8 blok bij discrete cosinustransformatie.

VP3

VP3 was oorspronkelijk een propriëtaire en gepatenteerde videocodec van On2 Technologies. In september 2001 is de codec vrijgegeven als vrije “open source” software waardoor het gebruikt kan worden door Theora en andere codecs die gebaseerd zijn op

VP3. In 2002 besloten On2 Technologies en Xiph.Org Foundation een nieuwe codec, Theora, te ontwikkelen, die op VP3 gebaseerd zou zijn. On2 Technologies bestempelde Theora ook als de opvolger van VP3.

Ondersteuning

Theora wordt ondersteund door allerlei programma's, zoals:

- Cortado (Java applet), eventueel in combinatie met ITheora (PHP-wrapper)
- FFmpeg
- MPlayer
- RealPlayer
- Helix Player
- VLC
- xine en mediaspelers die gebaseerd zijn op libxine, zoals Kaffeine
- Totem
- QuickTime 7
- Visonair.tv Player
- Miro Media Player (voorheen bekend als Democracy Player)

3.2.8 DV

Digitale Video (DV) is een digitaal video formaat voor opslag op tape ontwikkeld door Sony en werd in 1995 gelanceerd. Het is sindsdien uitgegroeid tot de standaard voor video productie voor amateuristisch of semi-professioneel gebruik. De DV-specificatie definieert zowel de codec als het tapeformaat. Er bestaat ook een verwant digitaal video formaat, miniDV, voor opslag op de kleinere tapes. DV levert een video kwaliteit die superieur is aan de analoge varianten zoals Video8, Hi8 en VHS-C.

DV gebruikt DCT intraframe compressie aan een vaste bitsnelheid van 25 Mbps, hetgeen samen met met de data voor het geluid , error detectie en error correctie resulteert in een bitsnelheid van ongeveer 36 Mbps. Tegen dezelfde bitsnelheid presteert DV iets beter dan de oudere MJPEG codec en is vergelijkbaar met intraframe MPEG-2 (waarbij enkel de I-frames intragecodeerd zijn). DCT-compressie is verlieslatend waardoor soms artefacten optreden rond kleine, complexe objecten zoals tekst. De DCT-transformatie is speciaal aangepast voor opslag op tape. Het beeld wordt verdeeld in macroblokken die elk bestaan uit 4 luminantie DCT blokken en 1 chrominantie DCT blok. Zes macroblokken, geselecteerd op posities die ver genoeg van elkaar liggen, worden gecodeerd in een vast aantal bits. Uiteindelijk wordt de informatie van elk gecomprimeerd macroblok zoveel mogelijk in één

sync-blok op de tape opgeslagen. Dit maakt het mogelijk om video op tape tegen hoge snelheid te doorzoeken dit in zowel voorwaartse richting als achterwaartse richting en om foute sync-blokken te corrigeren.

Het DV formaat gebruikt “L-size” cassettes, terwijl de MiniDV de zogenaamde “S-size” cassettes gebruikt. Beide cassettes bezitten een ingebed geheugen gaande van 4 Kbit voor MiniDV cassettes tot 16Kbit. Dit geheugen kan gebruikt worden om data allerhande op te slaan zoals een inhoudstafel, tijden en datums van de opnames en camera settings. Het is een EEPROM geheugen dat gebruik maakt van het PC protocol. Dit geheugen wordt wel zelden aangewend op gebruikersnivea. De meeste cassettes voor gebruikers bezitten zelfs geen chip, die de prijs van een cassette gevoelig verhoogt. De gebruikers apparatuur bezit meestal wel de nodige electronica om gegevens van de chip in te lezen en weg te schrijven, alhoewel het maar weinig wordt gebruikt.

Er bestaan verschillende varianten op de DV standaard. De meest gekende zijn Sony`s DVCAM en Panasonic`s DVCPRO voor professioneel gebruik. Sony`s gebruikersformaat Digital8 is een andere variant die gelijkend is op het DV formaat, maar aangepast voor opslag op Hi8 tape.

DVCAM van Sony is een professionele variant van de DV standaard en gebruikt dezelfde codec en cassettes als DV en MiniDV, maar transporteert de tape 33% sneller. Hierdoor is DVCAM veel naukeuriger te editeren. Een andere eigenschap van DVCAM is “locked” audio. ALS DV wordt gecopieerd kan na een aantal generaties van kopieën de audio niet meer gesynchroniseerd loopt met het beeld. Met DVCAM gebeurt dit niet.

Panasonic onhntwikkelde de DVCPRO codec om een betere lineaire editering te hebben van het DV formaat. De tape is naast een controlespoor voor het beter editeren voorzien van een longitudinale analoog audiospoor. De audio is beschikbaar in de 16 bit/48 kHz variant. DVCPRO gebruikt ook steeds 4:1:1 kleursubsampling (zelfs bij PAL). Op bitstream niveau is de standaard DVCPRO (DVCPRO25) identiek aan de standaard DV. DVCPRO werd door Panasonic aangeprezen als zijn DV variant voor professionele high-end applicaties. DVCPRO50 is in feite de combinatie van twee DVCPRO codecs in parallel. De bitsnelheid wordt dus verdubbeld tot 50Mbps en gebruikt 4:2:2 chroma subsampling in plaats van 4:1:1. De resulterende videokwaliteit is vergelijkbaar met deze van zijn rivaal, Digital Betacam. DVCPRO HD, ook wel gekend als DVCPRO100, gebruikt vier parallelle DVCPRO codecs, resulterend in een bitsnelheid van 100Mbps. DVCPRO HD maakt gebruik van 4:2:2 kleursampling.Deze codec is dus geschikt voor HD materiaal op te slaan op tape.

3.2.9 Betacam

Betacam is een familie van professionele videotape producten ontwikkeld door Sony. Betacam kan staan voor zowel de camcorder, de tape, de video recorder of het formaat. Het originele Betacam formaat werd gelanceerd in augustus 1982. Het is een analoog video formaat. De luminantie, Y, werd opgeslagen op één spoor en de chrominantie op een ander spoor als afwisselende segmenten van de R-Y en B-Y componenten. Dit leverde een kwaliteit van 300 lijnen horizontale luma resolutie en 120 lijnen chroma resolutie. Een nadeel van dit formaat waren de cassettes die slechts een opnametijd hadden van een half uur. In 1986 werd het Betacam SP formaaat ontwikkeld waarmee de horizontale resolutie werd verhoogd tot 340 lijnen. De kwaliteitsverbetering hiervan was gering, maar gecombineerd met een nieuwe cassette die 90 minuten kon opnemen, werd Betacam SP de industriestandaard voor de meeste TV-stations en high-end productiehuizen tot laat in de jaren `90.

Het digitale formaat, Digital Betacam (digibeta of d-beta), werd gelanceerd in 1993. Het verving het Betacam en Betacam SP formaat en de cassettes hadden een opnametijd van 40 minuten of 124 minuten. Het digitale Betacam formaat neemt een DCT-gecomprimeerd signaal op met 10 bit YUV 4:2:2 sampling in NTSC- (720x486) of PAL- (720x576) resolutie. Daarnaast worden nog eens vier kanalen opgenomen met 48 kHz 16 bit PCM audio. Een vijfde analoog audio spoor is ook beschikbaar voor cueing. Digitale Betacam gebruikt temporele compressie, waarbij een sequentie van I- en B-beelden wordt opgenomen. Het is een populair digitaal video formaat bij de zenders.

Betacam SX is een digitale versie van Betacam SP dat in 1996 werd geïntroduceerd als een goedkoper alternatief voor de digitale Betacam. Het slaat de video op gebruik makend van de MPEG-2 4:2:2 compressie, tesamen met 4 kanalen voor de audio. Dit resulteert in een betere chroma resolutie en laat bepaalde postproductieprocessen toe. De cassettes hebben een opnametijd van 62 of 194 minuten.

MPEG IMX is een ontwikkeling van het digitale Betacam formaat van 2001. Het gebruikt de MPEG compressie zoals Betacam SX, maar tegen een hogere bitsnelheid. De toegepaste compressie heeft drie formaten: 30 (6:1 compressie), 40 (4:1 compressie) of 50Mbit/s (3.3:1 compressie). Dit laat toe om de video op te slaan aan verschillende kwaliteit/opslag efficientie ratios. De video is opgenomen met het MPEG-2 4:2:2 profiel.

HDCAM, dat werd geïntroduceerd in 1997, is een High Definition-versie van het digitale Betacam formaat. Het gebruikt een 8-bit DCT compressie en het 3:1:1 profiel. De resolutie is

1440x1080 pixels waarmee het compatibel is met 1080i. De bitsnelheid is 144Mbit/s. Voor de audio worden vier kanalen gebruikt van 20 bit/48 kHz digitale audio. HDCAM SR, de opvolger in 2003 van HDCAM, kan opnemen in 10 bits 4:2:2 of 4:4:4 RGB tegen een bitsnelheid van 440 Mbit/s. Voor de compressie gebruikt HDCAM SR het nieuwe MPEG-4 Part 2 Studio profiel en het aantal audio-kanalen wordt verhoogd tot 12 24 bit/48kHz kanalen.

3.2.10 MP2

MPEG-1 Audio Layer II (MP2, of ook wel Musicam genoemd) is een audio codec gedefinieerd door de ISO/IEC 11172-3 standaard. Terwijl MP3 veel populairder is voor PC en internet applicaties, blijft MP2 de dominante standaard voor zenders.

De ontwikkeling van de MP2 standaard werd gestart eind jaren `80 door ISO`s Moving Picture Expert Group. MP2 is een psycho-acoestisch compressie algoritme. Dit wil zeggen dat het informatie verwijdert die voor het menselijk gehoor quasi niet waarneembaar is. Om te bepalen welke signalen niet waarneembaar zijn, wordt het audio signaal geanalyseerd volgens een psycho-akoestisch model, die de karaktereistieken aanneemt van het menselijk gehoor.

Uit studies hieromtrend is gebleken dat wanneer er een sterk signaal is op een bepaalde frequentie, de zwakkere signalen op naburige frequenties niet meer waarneembaar zijn. Hiervan maakt MP2 gretig gebruik. MP2 verdeeld het audio signaal over 32 frequentie subbanden. Wanneer de audio van een subband niet waarneembaar is, wordt deze subband weggelaten. MP3 b.v. verdeeld het audio signaal over 576 frequentie subbanden, waardoor MP3 een hogere frequentie resolutie heeft. MP3 gebruikt daarboven ook nog entropie codering, wat verklaart waarom MP3 lagere bitsnelheden nodig heeft dan MP2 om een vergelijkbare audiokwaliteit te hebben.

MP2 is minder rekenintensief dan MP3 wat de codec efficiënter maakt voor hoge kwaliteit percussieve geluiden (impusen), dankzij de goede tijdsdomein karakteristieken van zijn filterbank. Een bijkomend voordeel hiervan is dat MP2 beter bestand is tegen digitale transmissiefouten. Mede hierdoor wordt MP2 nog steeds gebruikt voor broadcast applicaties.

MP2 maakt deel uit van de DAB digitale radio en DVB digitale televisie standaarden. Ook de meeste DVD-spelers bezitten een MP2 decoder, waardoor in deze markt MP2 een concurrent is van Dolby Digital.

3.2.11 MP3

MPEG-1 Layer 3 is een manier om geluid te comprimeren en is dus een broncodering-techniek. De veel gebruikte afkorting is MP3. Het is een MPEG (Moving Picture Experts Group) standaard uit 1992, waarvan sinds 1994 implementaties bestaan.

Compressie

Met MP3 is het mogelijk de hoeveelheid opslagcapaciteit voor geluid sterk te verminderen. Dat gebeurt door geluidselementen, die voor mensen niet waarneembaar zijn, weg te laten. Een iets zachtere toon vlak naast een luide toon is bijvoorbeeld niet hoorbaar en kan dus weggelaten worden. Ook wordt er gebruik gemaakt van klassieke compressie waarbij informatie die zowel op het linker- als rechterkanaal voorkomt, slechts eenmaal wordt opgeslagen. Dit laatste noemt men "joint stereo".

MP3-compressie is lossy. Bij de compressie gaan dus gegevens verloren. Hierdoor kan de oorspronkelijke vorm niet meer volledig teruggewonnen worden maar enkel benaderd worden. Een MP3 muziekbestand kan bij het afspelen dus licht verschillen van het origineel (vlakker, bijgeluiden), hoewel dat bij blinde luistertests en bij hoge bitrates met een goede encoder (zoals LAME) nauwelijks aantoonbaar was. Dat er sprake is van verlies in kwaliteit zal in veel gevallen pas merkbaar zijn na een aaneenschakeling van verschillende encodeer-decodeer stappen.

Het is wel mogelijk om MP3-bestanden zonder verlies aan informatie (lossless) te bewerken. Een MP3-bestand bestaat uit kleine pakketjes van een fractie van een seconde. Bij elk pakketje wordt het volume aangegeven. Dat volume kan achteraf gewijzigd worden zonder de code van het gecomprimeerde geluid te wijzigen. Ook kunnen pakketjes worden verwijderd.

Door de meeste gebruikers wordt de kwaliteit van MP3's met een bitrate van 192 kbps of hoger (kilobits per seconde) vrij goed bevonden. Bij een bitrate van 320 kbps verschilt de kwaliteit niet hoorbaar van een CD. Veel gangbare bitrates zijn een veelvoud van 32 of 64 kbps: 128, 160, 192, 256 of 320 kbps. Een MP3-bestand met een bitrate van 128 kbps is elf keer zo klein als hetzelfde geluidbestand in WAV-formaat. Eén minuut geluid heeft dan een grootte van ca. 1 MegaByte. Bij hogere bitrates neemt de bestandsgrootte bij dezelfde sample rate (en 16 bits) evenredig toe.

Verbeterde versies

Thomson Multimedia en het Fraunhofer instituut hebben na het grote succes van het MP3-formaat ook verschillende verbeterde versies ontwikkeld: mp3PRO en AAC. mp3PRO zou met de helft van de bitrate dezelfde kwaliteit bieden als MP3.

Patent

De compressie- en decompressiealgoritmen van MP3 zijn gepatenteerd door de eigenaar, het Fraunhofer instituut, en dus niet vrij beschikbaar voor commerciële producten of commercieel gebruik van de technologie. Persoonlijk gebruik van de MP3-software is toegestaan. “Open source” encoders en decoders worden toegelaten. Een patentvrij alternatief voor MP3 is Ogg Vorbis. Andere alternatieven, zoals SHN en FLAC, zijn gratis voor niet-commercieel gebruik.

3.2.12 AAC/MPEG-2 Part 7/MPEG-4 Part 3

Advanced Audio Coding (AAC) is een MPEG-2/MPEG-4 tegenhanger geworden van het populaire MP3-audioformaat. Anders dan MP3 ondersteunt AAC multi-kanaals audio tot maximaal 48 full frequency channels. Een 192 kbps MP3-geluidsfragment zal ongeveer dezelfde kwaliteit hebben als een 128 kbps AAC bestand. Het AAC bestand is dus kleiner en heeft een efficiëntere datacompressie (40% kleiner) en een betere geluidskwaliteit dan MP3. AAC is op dit ogenblik het standaard audioformaat voor Apple’s iPhone, iPod en iTunes, voor Sony’s PlayStation3, hun laatste generatie Walkman en Walkman Phone, voor Nintendo’s Wii en de MPEG-4 videostandaard.

AAC+ of HE-AAC

AAC+ combineert drie technieken: AAC, Spectral band replication (SPR, reproduceert hoge tonen) en Parametric Stereo (PS, combineert 2 monostreams en maakt er een stereosignaal van terwijl er slechts 2-3kbps extra informatie wordt gebruikt). De codec is vooral geschikt voor lage bitrates. Op bitrates van 32-64 kbps heeft AAC+ doorgaans de beste geluidskwaliteit, in vergelijking met andere codecs. Bij hogere bitrates is AAC+ nog steeds beter dan MP3, een AAC+ van 80/96 kbps klinkt ongeveer zoals een 128/140 kbps MP3, maar is even groot als de 96 kbps MP3. Vanaf 128 kbps en meer zal AAC+ hetzelfde klinken als MP3. AAC+ wordt vooral gebruikt voor internetradio en mediaspelers waar de opslagcapaciteit niet al te hoog is.

Ondersteuning

AAC+ is nog niet heel goed ondersteund. Winamp en Foobar2000 zijn de voornaamste mediaspelers die AAC+ ondersteunen. Ook iTunes ondersteunt AAC+.

Andere namen

Er zijn meerdere namen voor AAC+, waarvan AAC+ en HE-AAC (HE = High Efficiency) meest gebruikt worden. Andere mogelijke namen zijn: aacPlus, aacPlusV2, eAAC+, en NeroDigital Audio.

3.2.13 FLAC

FLAC is een afkorting van Free Lossless Audio Codec. Het wordt gebruikt wanneer men audio wil digitaliseren zonder verlies (lossless). FLAC is het snelste en best ondersteunde lossless-formaat. Bovendien is FLAC gratis en een “open source” project. FLAC haalt een compressie van 30%-50%, waar MP3 80% haalt. MP3 doet dit echter met enig kwaliteitsverlies.

FLAC wordt ondersteund door zowat alle huidige besturingssystemen. Zowel op Windows als op OSX of op een Linux distributie bestaan spelers die FLAC ondersteunen. FLAC wordt ook ondersteund door de meeste bekende mediaspelers, zoals Winamp, XMMS, Media Player Classic, Foobar2000 en Windows Media Player, na installatie van K-lite mega codec pack, VLC media player en Songbird.

Techniek

Een audiobron zoals een cd wordt, door gebruik van FLAC als compressie, 40 à 50% kleiner dan het originele exemplaar (47% volgens de uitgevers) zonder verlies van informatie. Een FLAC audio-bestand kan tot maximaal 8 kanalen bevatten. Dat wil zeggen dat je zowel mono- als surroundgeluid kan opslaan in FLAC. De sampling rate ligt tussen 1 en 1.048.570 Hz gaan, in stappen van 1 Hz.

3.2.14 Ogg Vorbis

Ogg Vorbis is ook een “open source” compressiemethode voor geluidsbestanden. In tegenstelling tot bijvoorbeeld de audio-indelingen MP3 en Microsofts WMA is het vrij van patenten. Het feit dat er enkel gepatenteerde audio-indelingen bestonden, was voor de ontwikkelaars van “open source” software aanleiding om een nieuwe multimedia-indeling te definiëren, waarin geen enkel patent zou voorkomen.

Een bijkomend voordeel bestond erin dat men van de nieuwste inzichten gebruik kon maken, waardoor de kwaliteit ten opzichte van de oudere MP3-indeling een stuk hoger zou zijn. “Ogg” staat voor het algemene formaat dat op zich verschillende componenten kan

omvatten, zoals Vorbis. Vorbis is het audiogedeelte. Ogg Vorbis is een compressietechniek die weinig relevante geluidsinformatie wegfiltert. Net zoals bij MP3 gaat hierbij dus een deel van de geluidsinformatie verloren, maar door gebruik te maken van de fysiologische en psychologische kenmerken van ons gehoor is dat nauwelijks hoorbaar.

Op de website van Ogg Vorbis wordt nauwkeurig verslag gedaan van alle vorderingen. Zoals alle “open source” projecten is Ogg Vorbis immers een werk in uitvoering waarbij men steeds inzicht krijgt in de laatste ontwikkelingen.

De meest recente versie 1.0.1 dateert van 2004 en bestaat zowel in een Windows-, Macintosh- als in een Linux-uitvoering. Zelfs voor het inmiddels verdwenen multimediabesturingssysteem BeOS bestond goede software. Het is die platformonafhankelijkheid die veel ontwikkelaars, o.a. van browsers en games, aantrekt. Zo wordt in het computerspel Unreal Tournament gebruik gemaakt van .ogg-geluidsbestanden. Ook de HTML5 werkgroep binnen W3C bekijkt Ogg Vorbis terdege om eventueel als standaard audiotag te embedden in HTML5 compliant browsers. De financiële eisen van licentiehouders van commerciële audioformaten blijven immers onvoorspelbaar.

3.2.15 AC-3/Dolby Digital

Dolby Digital is de verzamelnaam van een batterij lossy audio compressietechnieken die worden ontwikkeld door Dolby Labatories. AC-3 is daarvan de meest gebruikte techniek die tot 6 (genaamd 5.1) discrete audiokanalen kan bevatten. Het ondersteunt een sample rate van 48 kHz. AC-3/Dolby Digital wordt voornamelijk gebruikt in digitale cinema en op DVD’s. De meer geavanceerde codecs Dolby Digital Plus en Dolby TrueHD worden ook ondersteund op de nieuwe DVD-formaten Blu-Ray en HD-DVD.

3.2.16 TTA

Het True Audio (TTA) formaat is vrije software die audio kan comprimeren. Na het expanderen, ontstaat exact hetzelfde signaal. Het gaat dus om een verliesvrije compressietechniek. Tijdens de compressie wordt gebruikt gemaakt van een prognose van de vorm van het signaal. Het uiteindelijke bestand is daardoor ongeveer 30% tot 70% kleiner. Net als bij MP3 kan informatie over het geluidsbestand in ID3v1 en ID3v2 tags opgeslagen worden.

3.2.17 Windows Media Audio

WMA is een audiocompressievorm die door Microsoft ontwikkeld is en standaard met Windows Media Player wordt meegeleverd. Het is een gesloten formaat. Aangezien Windows een groot marktaandeel heeft, is deze codec snel de facto een standaard geworden. WMA staat voor Windows Media Audio en komt meestal samen met het WMV-formaat voor, dat de video-tegenhanger is van het WMA-formaat. WMA is vooral bedoeld voor het streamen op lage bandbreedtes. Een groot verschil met andere formaten zoals MP3 en OGG Vorbis is het mogelijke gebruik van DRM, een controversiële techniek die gebruikers beperkt in de mogelijkheden bestanden te openen, te kopiëren etc. als de auteur van die bestanden beperkingen oplegt.

De werking van WMA is grotendeels te vergelijken met die van het MP3-formaat. Het maakt ook gebruik van psycho-akoestische schema's maar het filtert alleen het geluid weg waarvan de frequentie meer dan 20 kHz en minder dan 20 Hz bedraagt. WMA-bestanden zijn echter kleiner dan MP3-bestanden, wat dan weer een groot voordeel is. Er kunnen zo immers meer muzieknummers op een walkman (vanaf ca. 2005 kunnen veel MP3-walkmans ook WMA-bestanden afspelen). WMA is geen ideale standaard voor algemene internettoepassingen. Het afspelen van dit formaat is namelijk enkel mogelijk in besturingssystemen en spelers die door Microsoft ondersteund worden.

3.2.18 JPEG

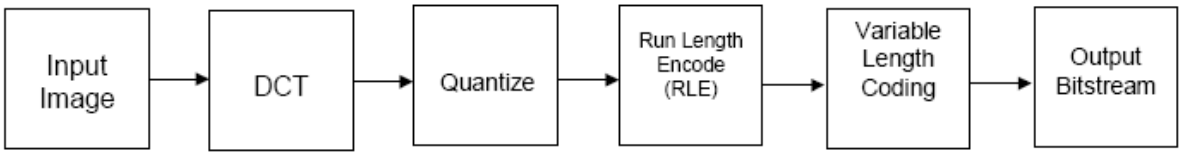
JPEG (Joint Photographic Experts Group), ontwikkeld door ISO/ITU, is de eerste algemeen verspreide compressiestandaard voor afbeeldingen. Het werd specifiek ontworpen om digitale beelden te comprimeren en is op dit moment de “de facto” standaard voor afbeeldingen op het internet en als opslagformaat voor digitale camera’s. De kwaliteit van de compressie hangt grotendeels af van de eigenlijke inhoud van de afbeelding: zijn er al dan niet veel details of hoog frequente componenten in de afbeelding aanwezig. Een compressieverhouding van 10:1 kan typisch gehaald worden zonder dat er aantoonbaar kwaliteitsverlies optreedt. Eens men hogere compressieratio’s gebruikt, zal men zeker kwaliteitsverlies in de afbeelding opmerken aan de hand van opduikende blokartefacten en wazige contouren.

De JPEG-bestandsindeling kent diverse compressiemogelijkheden. Hoe hoger de compressie hoe kleiner het bestand en hoe geringer de beeldkwaliteit. Het kwaliteitsverlies van JPEG valt bij foto’s niet erg op maar wel bij bijvoorbeeld grafieken, lijnen of letters. Voor dit soort afbeeldingen is de GIF- of PNG-compressie beter geschikt, ofwel de nieuwe RAW-

methode, waar de opslag plaatsvindt zonder compressie en het beeld later via geschikte software bewerkt kan worden. Daarom wordt JPEG doorgaans gebruikt voor foto's, bijvoorbeeld met digitale camera's gemaakt.

Werking

De JPEG-indeling is complex. In tegenstelling tot indelingen als PNG of GIF wordt niet van één enkel mechanisme gebruikgemaakt maar er wordt een groot aantal stappen na elkaar toegepast om tot het uiteindelijke JPEG-bestand te komen.



De figuur hierboven toont de hoofdmechanismes die nodig zijn om tot een JPEG-afbeelding te komen:

- **Blokgebaseerde verwerking**: elke afbeelding wordt onderverdeeld in pixelblokken (typisch 8x8) zodat de verwerking van de volledige afbeelding kan opgesplitst worden tot op dit blokniveau.

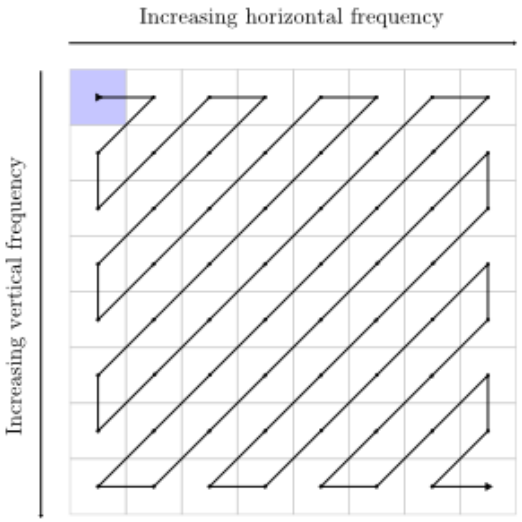
- **Intra-frame codering**: enkel spatiale overbodigheden die binnen de afbeelding zelf bestaan worden uitgebuit door ze door een reeks filters (transformatie, quantisatie & entropie codering) te sturen.

- **8x8 discrete cosinus transformatie (DCT)**: elk 8x8 blok van spatiale pixelwaarden wordt vertaald naar het frequentiedomein door middel van een DCT waardoor er 64 frequentiecomponenten onstaan. In dat nieuwe blokje is de pixel linksboven de gemiddelde kleur van het plaatje. Als je nu naar rechts of naar onder beweegt, dan bepalen die pixels details van steeds kleinere gebieden. De discrete cosinustransformatie is wiskundig gezien exact omkeerbaar maar in de praktijk gaat er wegens de beperkte rekenkundige precisie in de computer wat informatie verloren.

- **Perceptuele quantisatie**: herschaal de bitallocatie van deze verschillende frequentiecomponenten zodanig dat er veel nulcoëfficiënten ontstaan. De pixels rechtsonder in het getransformeerde blokje bepalen immers de fijnste details. Als we die weggooien, vallen alleen enkele fijne details weg. Dat is ook wat men bij JPEG doet. In de JPEG-standaard zijn een aantal quantisatie-matrices opgenomen. Afhankelijk van de kwaliteit die de gebruiker kiest, wordt een andere matrix gekozen.

In deze 8x8 matrix staat voor iedere pixel een getal. Voor alle pixels in het getransformeerde blokje wordt de pixelwaarde gedeeld door het getal in deze quantisatiematrix. Bij hoge kwaliteit zullen veel pixels door 1 gedeeld worden. Naarmate de kwaliteit teruggeschroefd wordt, zullen de pixels rechtsonder door grotere getallen (vb. 128) gedeeld worden. Omdat hier met gehele getallen gerekend wordt, wordt informatie weggegooid. Door herhaaldelijk te delen, verkrijgt men rechtsonder in de matrix veel lage pixelwaarden (liefst nul).

- **Run-length codering**: representeer deze gequantitiseerde frequentiecomponenten op zo'n manier dat de niet-nulcoëfficiënten niveaus de nulcoëfficiënten voorafgaan. Vervolgens worden de pixels dus gelineariseerd tot één lange rij van 64 pixels. Dat gebeurt in een speciale volgorde: de volgende tabel laat zien welke pixel op welke plaats in de rij komt:



- Variabele lengte (Huffman) codering: in de laatste stap wordt het getransformeerde blokje van 8x8 pixels met een Huffmancodering bewerkt. Het resultaat daarvan wordt in het JPEG-bestand opgeslagen.

Gegevens in het jpg-bestand

Een JPG-bestand kan verschillende gegevens bevatten. Dit worden APP markers genoemd:

- Comment: vrij in te vullen commentaar.
- EXIF: gegevens van de digitale camera of scanner.
- IPTC: gestandaardiseerde indeling voor gegevens van de afbeelding.
- Andere niet-standaard gegevens.

Verliesvrij bewerken

Omdat de JPEG-compressie op dezelfde wijze in horizontale en in vertikale richting werkt, is het mogelijk om een afbeelding, die al als een JPEG-bestand is opgeslagen, verliesvrij 90 graden te roteren (en ook 180° en 270°). Ook verliesvrij spiegelen (horizontaal en vertikaal) is mogelijk.

3.2.19 JPEG-LS

JPEG-LS is de huidige ISO/ITU-standaard voor lossless of bijna-lossless compressie en maakt deel uit van een meer uitgebreide ISO-standaard voor het beter comprimeren van medische beelden.

Intro

Lossless compressie is van zeer groot belang voor afbeeldingen die kritische informatie bevatten. Dat kan het geval zijn bij beelden uit de medische wereld of bij afbeeldingen die moeilijk en duur te produceren zijn. Eerdere lossless compressiemethodes zoals GIF en PNG zijn enkel efficiënt bij de compressie van afbeeldingen die slechts een beperkte hoeveelheid kleuren bevatten. JPEG-LS maakt het mogelijk om ook afbeeldingen met meerdere kleuren verliesvrij te comprimeren. Ook de nieuwe JPEG 2000-standaard voorziet een lossless mode. Het algoritme dat daarbij gebruikt wordt, is echter veel complexer dan het JPEG-LS-algoritme.

Lossless: LOCO-I

Het kernalgoritme van JPEG-LS draagt de naam LOw COmplexity LOssless COmpression for Images (LOCO-I) en werd ontwikkeld door Hewlett-Packard. Het uitgangspunt van LOCO-I is dat de vereenvoudiging van een algoritme vaak tot betere resultaten leidt dan de uitbreiding van het algoritme. Een uitbreiding maakt het algoritme immers complexer en zorgt vaak voor slechts een kleine compressietoename.

Context modeling

LOCO-I maakt gebruik van een concept dat “context modeling” wordt genoemd. Tijdens het compressieproces wordt immers berekend wat de voorwaardelijke kans is dat een bepaalde pixel zal volgen op een andere pixel in de afbeelding. Deze extra informatie wordt de context genoemd en zal mee als input worden gebruikt voor de compressie. Zo wordt het mogelijk om een compressie uit te voeren die minder bits nodig heeft dan een entropie van de 0-de orde.

LOCO-I maakt gebruik van een contextmodel dat wordt weergegeven a.d.h.v. het volgende patroon:

C	**A**	**d**
B	**X**	**...**

Bij het scannen van het raster zullen de contextpixels a,b,c en d eerste gescand worden, vóór x dus. Men spreekt in dat verband ook wel eens van een causale context.

Voorspelling

In deze stap wordt de waarde van de volgende sample x' voorspeld. Dat gebeurt door het uitvoeren van steeds dezelfde primitieve test. Het zo eenvoudig mogelijk houden van deze test is één van de sleutels voor het eenvoudig compressiealgoritme van LOCO-I. De test wordt gegeven door

$$\hat{a}' = \begin{cases} min(a,b) & c \geq max(a,b) \\ max(a,b) & c \leq min(a,b) \\ a+b-c & anders \end{cases}$$

Door deze predictor worden vertikale en horizontale kleurovergangen gedetecteerd. Wanneer er links naast de huidige pixel een vertikale overgang is, dan wordt de output a. Is er een horizontale rand boven de huidige pixel, dan wordt de output b. Werd er geen duidelijke overgang gedetecteerd, dan wordt a + b − c teruggegeven.

Contextbepaling

Aangezien het slechts om een voorspelling gaat, zal er steeds sprake zijn van een mogelijke fout. Deze wordt de voorspellingsfout of residu genoemd. Het ‘context model’ dat dit residu bepaalt, wordt aangeduid door de contextvector Q=(q1,q2,q3) waarbij q1, q2 en q3 de lokale overgangen (verschillen) of gelijkenissen voorstellen: q1=d – b, q2=b – c, q3=c − a.

Residu codering

Ten slotte wordt het residu gecodeerd met behulp van Golomb codes, die ideaal zijn voor de codering van twijzijdige geometrische verdelingen aangezien kan aangetoond worden dat het residu tweezijdig geometrisch verdeeld is.

Bijna-lossless mode

JPEG-LS voorziet ook een bijna-lossless mode waarin de gereconstrueerde samples een maximale afwijking δ hebben ten opzichte van het originele beeld. De lossless mode kan eigenlijk beschouwd worden als een speciaal geval van bijna-lossless, waarbij δ = 0.

3.2.20 JPEG-2000

JPEG 2000 is een nieuwe standaard voor compressie van digitale beelden, met name van 'continue-tint'-grijswaarden/kleurenbeelden en binaire beelden. JPEG is het acroniem voor Joint Pictures Experts Group.

Deze standaard is het resultaat van een gezamenlijke inspanning van de Internationale Standaardisatie Organisatie (ISO), de Internationale Elektrotechnische Commissie (IEC) en de Internationale Telecommunicatie Unie (ITU-T). Het kerncodeersysteem van de JPEG 2000 standaard is officieel geregistreerd als ISO/IEC 15444-1/ITU-T Rec. T.800. Deze standaard levert voor continue-tint-beelden excellente compressieprestaties in termen van bitdebiet/distorsie-gedrag en overtreft, in het bijzonder bij lagere bitdebieten, de prestaties van zijn voorganger JPEG. De focus bij JPEG 2000 ligt op kwaliteit en functionaliteit en hiervoor benut men nieuwe technologieën.

De discrete cosinustransformatie (DCT) is bij JPEG 2000 vervangen door de wavelet transformatie, een overgang van een lokale, blokgebaseerde transformatie naar een globale beeldtransformatie. Hierdoor worden de storende blokartefacten bij lage bitdebieten vermeden ten koste van meer uitgesmeerde beelden. De wavelet transformatie leent zich bovendien erg goed voor schaalbare of ingebedde codering, waarbij de prestaties die geleverd worden door de hiërarchische en progressieve DCT-mode van de oude JPEG-standaard, overschaduwd worden.

Samengevat ondersteunt of levert JPEG 2000 superieure codeerprestaties bij lage bitdebieten, de compressie van continue-tint en binaire beelden, een grote dynamische range van de pixels, grote beelden en een groot aantal beeldcomponenten, verliesloze en verlieshebbende compressie, bitdebietoptimalisatie, progressieve transmissie in termen van kwaliteit en resolutie, interesseregiocodering, willekeurige toegang en bewerkingen in het gecomprimeerde domein, objectgebaseerde functionaliteit, robuustheid tegen bitfouten (foutresistentie), mogelijkheid tot sequentiële codering, een fileformaat, (JPX) en de mogelijkheid tot beeldbeveiliging.

De penetratie van JPEG 2000 is nog niet heel groot in vergelijking met zijn voorganger JPEG. Dat is enerzijds te wijten aan licentieclaims met betrekking tot de oorspronkelijke JPEG-standaard, wat tot heel wat onzekerheid op de markt leidde, en anderzijds aan de hogere complexiteit. Ondertussen zijn deze problemen van de baan en zien we dat zowel voor low-end als high-end devices JPEG2000 zijn intrede begint te maken. JPEG2000 is wel doorgebroken op high-end markten zoals videobewaking en medische beeldvorming waar zowel kwaliteit als schaalbaarheid van belang zijn. Ook voor digitale cinema werd JPEG 2000 als codeerstandaard geselecteerd.

3.2.21 GIF

GIF is een bestandsindeling voor het opslaan van afbeeldingen in digitale vorm.

GIF is de afkorting van Graphics Interchange Format, een grafische bestandsindeling met pixels. GIF ondersteunt kleuren, verschillende resoluties, animatie en een transparante achtergrond. Het aantal kleuren in een GIF-bestand is beperkt tot maximaal 256 (door het gebruik van 8 bits), die ieder uit 262.144 verschillende tinten gekozen kunnen worden.

Compressie vindt plaats op basis van de verdeling en het aantal kleuren in horizontale richting. Indien het een afbeelding met weinig kleuren en met herhalende patronen betreft, is goede compressie mogelijk en is de bestandsgrootte erg klein. Zijn er veel kleuren of is er dithering toegepast dan neemt de bestandsgrootte toe en zijn bestandsformaten als JPEG of PNG met 24 bits per pixel veelal een betere optie.

Geschiedenis

De GIF-bestandsindeling is populair geworden/gemaakt door Compuserve in de jaren tachtig, vanwege de mogelijkheid om grafische informatie over netwerken te versturen. In de jaren negentig is de GIF-indeling overgenomen door ontwikkelaars van het Internet om websites op te luisteren. Tegenwoordig komt het bestand op veel websites en in veel bewegende plaatjes voor.

Voor de compressie wordt gebruikt gemaakt van de LZW compressietechnologie. Deze technologie is gepatenteerd door Unisys. Hierdoor moesten toeslagen betaald worden voor het gebruik van applicaties die deze compressie toepassen (m.n. de bewerkingsprogramma's). Dat was een van belangrijkste redenen voor de ontwikkeling van een rechtenvrije grafische bestandsindeling zoals PNG.

Het Verenigde Staten LZW patent (No. 4,558,302) is verjaard op 20 juni 2003. Het Canadese patent liep af op 7 juli 2004, de patenten voor Engeland, Frankrijk, Duitsland en Italië waren geldig tot 18 juni 2004 en het Japanse patent tot 20 juni 2004.

Volgens een onderzoek van de Free Software Foundation is het laatste patent (van IBM) op 11 augustus 2006 verlopen.

Animated GIF

GIF maakt het mogelijk om verschillende beelden na elkaar op te slaan in hetzelfde bestand, waardoor er een klein tekenfilmpje wordt vertoond.

Kleurreductie en dithering

Doordat een GIF-bestand maximaal 256 kleuren kan hebben, is het niet erg geschikt voor (kleuren)foto's. Om een foto optimaal weer te geven, zal het palet van 256 beschikbare kleuren zo goed mogelijk verdeeld moeten worden. Ten eerste moeten uit de mogelijke 262144 kleuren de benodigde kleuren goed gekozen worden. In veel beeldbewerkingsprogramma's kan de gebruiker kiezen uit een aantal vaste paletten of een "optimaal palet". Algoritmes die een optimaal palet berekenen gaan meestal uit van een driedimensionaal histogram van de in het origineel gebruikte kleuren en splitsen dit op in deelruimtes. Elke deelruimte wordt vervolgens ook weer gesplitst tot het gewenste aantal deelkleuren bereikt is. Dit leidt automatisch tot de "gemiddeld" meest gebruikte kleuren. Bij een portret zullen bijvoorbeeld meer huidtonen gekozen moeten worden, een landschap (zie het voorbeeld hieronder) bestaat voornamelijk uit blauw- en grijstonen.

Daarnaast kan de schijnbare kleurfout nog verder beperkt worden door een techniek die met de Engelse term 'ditheren' of 'error diffusion' aangeduid wordt. Hierbij wordt de kleur van een enkele pixel niet alleen bepaald door de waarde van de originele pixel maar ook door de afwijking in kleur van de omliggende pixels. Op die manier ontstaat een ietwat korrelig patroon dat gemiddeld exact de juiste kleuren heeft. Als men de afbeelding op een afstand bekijkt zodat individuele pixels niet meer zichtbaar zijn, ziet men nauwelijks dat het aantal kleuren beperkt is.

Elk programma zal op een andere manier 'ditheren'. Het origineel bevat zo'n 10.000 verschillende kleuren.

3.2.22 PNG

PNG is een bestandsformaat voor afbeeldingen met verliesloze compressie. De afkorting staat voor Portable Network Graphic, maar soms wordt ook het recursieve backroniem PNG's not GIF gebruikt.

Het PNG-formaat, dat sinds 1995 bestaat, is in het leven geroepen om een patentvrij alternatief te bieden voor het populaire GIF-formaat. Het GIF-formaat maakt namelijk gebruik van de gepatenteerde LZW-compressie, waarover het technologiebedrijf Unisys toentertijd heeft besloten dat er betaald moet worden voor licenties om LZW in programma's te mogen gebruiken. Dit is ook de reden waarom het PNG-formaat vaak wordt gebruikt in open-source-programma's.

In vergelijking met andere formaten zoals BMP en TGA nemen PNG-afbeeldingen relatief weinig ruimte in maar ze hebben dezelfde kwaliteit. Andere voordelen van het formaat zijn de mogelijkheden tot gedeeltelijke transparantie en de ondersteuning van ruim zestien miljoen kleuren, terwijl het GIF-formaat nog steeds beperkt is tot maximum 256 kleuren. Een PNG-afbeelding kan echter ook, net als een GIF-afbeelding, een "palet" hebben en dus maximum 256 gebruiken. Door het kleiner aantal bits per kleur, verkleint de bestandsgrootte hierdoor aanzienlijk. Op deze manier is PNG zowel voor verliesloze opslag van afbeeldingen als voor het besparen van geheugenruimte voor simpele afbeeldingen zeer geschikt. Voor foto's kan het JPEG-formaat, dat kleinere bestanden oplevert maar met verlies van beeldinformatie, een alternatief blijven bieden.

In een PNG-afbeelding kan elke pixel niet alleen een rood-, groen- en blauwwaarde bevatten maar ook een transparantie (alpha-waarde). Het gevolg hiervan is dat elke pixel een bepaalde hoeveelheid transparantie kan hebben, bijvoorbeeld helemaal doorzichtig of gedeeltelijk doorzichtig met wat rood eroverheen.

Vroeger kon PNG geanimeerde beelden niet ondersteunen. Sinds kort bestaat er het zogenaamde APNG of Animated PNG, een PNG-afbeelding die animaties ondersteunt. Reclamemakers ontdekten de animatiemogelijkheden van GIF op het internet op een moment waarop PNG geïntroduceerd werd. Hierdoor vertraagde de opkomst van PNG en vooral het verdwijnen van GIF. Tegenwoordig wordt voor bewegende advertenties vaak GIF of Flash gebruikt.

Een ander verwant formaat is JNG, dat JPEG-compressie in een PNG-achtig formaat biedt. Het is vooral ontworpen als mogelijke combinatie met MNG.

Doordat de veelgebruikte internetbrowser Microsoft Internet Explorer tot en met versie 6.0 het PNG-formaat qua transparantie niet volledig ondersteunt, is het PNG-formaat nog niet zo populair als GIF. In ieder geval stijgt het gebruik van PNG en daalt dat van GIF. Internet Explorer 7 beschikt wel over correcte PNG-ondersteuning. Zo goed als elk modern beeldverwerkingsprogramma ondersteunt het PNG-formaat.

3.2.23 TIFF

TIFF staat voor Tagged Image File Format. Op dit moment is TIFF, naast JPEG en PNG, het formaat voor het opslaan van beelden. Het is tevens het meest universele en ondersteunde formaat voor alle platformen, MAC, Windows en UNIX.

TIFF was oorspronkelijk mid jaren `80 ontworpen om een gemeenschappelijk beeldformaat te hebben voor de desktop scanners. In het begin was TIFF enkel een binair beeldformaat, met enkel twee mogelijke waarden voor een pixel. Naarmate de scanners geavanceerder werden en opslagruimte minder schaars, begon TIFF ook grijswaarden te ondersteunen en uiteindelijk ook kleurwaarden. TIFF ondersteunt nu de meeste kleurenruimtes zoals RGB, CMYK, YcbCr, ...

TIFF laat zich onderscheiden van de meeste andere beeldcompressieformaten doordat de beeldheader flexibel is en zelf te definiëren. De header kan dus zelf samengesteld worden door een set van informatievelden of tags. Deze tags kunnen de meest elementaire informatie bevatten zoals beeldgrootte of bitvolgorde, maar ze kunnen ook b.v. de rechten beschrijven. Er bestaat zelfs de mogelijkheid om `private` tags te gebruiken die de eigen applicatie specifieke informatie bevatten. Het voordeel hiervan is dat de data kan vergezeld worden van gelijk welke informatie. TIFF kan uiteindelijk gezien worden als een containerformaat voor beelden. Zij biedt ondersteuning voor multipage, meerdere beelden binnen één file en multilayer, meerdere lagen binnen een beeld.

Via twee tags kan ook de gebruikte compressie en kleurenruimte worden gedefinieerd. TIFF laat dus toe om gelijk welke compressie te gebruiken in combinatie met gelijk welke kleurenruimte, als de portabiliteit van het bestand buiten beschouwing wordt gelaten. TIFF kan gebruikt worden met of zonder compressie. Zo wordt G3-compressie gebruikt als de standaard voor fax en multi-page bestanden. Optioneel kan ook de LZW verliesloze compressie toegepast worden. Verliesloze compressie is dus mogelijk met het TIFF formaat. De compressie zal herhaalde identieke strings detecteren en vervangt deze instanties door één instantie op zo`n manier dat het zonder verlies terug kan gedecodeerd worden. Het

gebruik van deze compressie veroorzaakt wel een vertraging in het openen en opslaan van de bestanden. LZW is het meest effectief wanneer solid indexed colors worden gecomprimeerd en is minder effectief voor 24bit coninuous fotoformaten. LZW is effectiever voor grayscale beelden dan kleurenbeelden. 48bit beelden leveren nauwelijks een compressie op. Zoals eerder vemeld ondersteunt TIFF beelden opgebouwd uit verschillende pagina`s of lagen, waardoor een TIFF bestand bijvoorbeeld een vector-gebaseerde clipping path kan bevatten.

Een ander krachtig mechanisme van TIFF is zijn ondersteuning van verschillende datatypes, gaande van signed of unsigned integers, floating point waarden tot zelfs complexe datastructuren. De combinatie hiervan met de mogelijkheid om verschillende beeldkanalen op te slaan, maakt van TIFF een heel handig formaat voor wetenschappelijke data. Zonder compressie wordt TIFF vooral gebruikt voor het archiveren van beelden waar kwaliteit belangrijk is.

Deze flexibiliteit heeft ook zijn nadelen. Nieuwe types zijn gemakkelijk te vormen, maar dit kan op zijn beurt weer incompabiliteit veroorzaken. Dit kan echter vermeden worden door de standaard TIFF types te gebruiken die door de meeste applicaties worden ondersteund. Een ander nadeel van TIFF is de beperking van de beeldgrootte die maximum 4 gigabytes groot is. Op dit moment is er wel een initiatief om deze beperking weg te werken, het BigTIFF bestandsformaat als opvolger van het TIFF bestandsformaat.

3.3 Fysieke containers

3.3.1 WAV

WAV, of Waveform Audio Format, is een audio bestandsformaatstandaard voor de opslag van audio op PC's. Het slaat de audiodata ruw op. Door het verliesloze karakter van ruwe audio kunnen deze WAV-bestanden echter heel groot worden.

Een WAV-bestand wordt opgebouwd uit zogenaamde chuncks. Deze chuncks geven informatie over het geluid of bevatten het geluid zelf. Naast deze chuncks bevat een WAV-bestand ook een header met onder ander informatie over de gebruikte formaatstructuur voor het bestand.

De maximale grootte van een WAV-bestand bedraagt 4GB wat overeenkomt met ongeveer 405 minuten geluid in CD-kwaliteit (44.1kHz, 16 bit, stereo) en 62 minuten in DVD-Audio-kwaliteit (tot 192kHz, tot 24 bit, stereo). Om deze beperkingen weg te werken, werd later het W64-formaat ontworpen dat de grootte van het bestand in 64 bits in plaats van in 32 bits beschrijft. De EBU heeft om dezelfde reden het RF64-formaat ontwikkeld. Dat formaat voegt verder ook nog onder andere de ondersteuning toe voor maximaal 18 surround kanalen. Naast ruwe audio ondersteunt de WAV-container ook andere codecs zoals GSM, ADPCM en MPEG Layer-3.

3.3.2 AIFF

AIFF, of Audio Interchange File Format, is de Apple Macintosh-tegenhanger van WAV. Dat formaat komt grotendeels overeen met het WAV-formaat van Microsoft. Het grote verschil bestaat erin dat waar bij WAV de samples in een little-endian-byte-volgorde worden opgeslagen, dit bij AIFF in big-endian-byte-volgorde wordt gedaan. Sinds Mac OS X heeft Apple echter een nieuw type AIFF gecreëerd dat in little-endian-byte-volgorde wordt opgeslagen. Dit door de overgang naar Intel-processoren die little-endian-byte-volgorde gebruiken.

AIFF is ook opgebouwd uit een header en zogenaamde chuncks die zowel de informatie over het geluid als het geluid zelf kunnen bevatten.

3.3.3 XMF

XMF, of eXtensible Music Format, is een familie van muziekgerelateerde formaten, ontworpen door de MIDI Manufacturer's Association. XMF heeft tot doel één of meerdere bestanden, in bestaande formaten zoals MIDI en WAV, samen te voegen.

XMF bestaat uit twee delen: het XMF Meta-File Format en een reeks XMF File Types, die gebruik maken van het XMF Meta-File Format. Tot dusver zijn XMF Type 0, XMF Type 1 en Mobile XMF gedefinieerd. Deze zijn echter allemaal op MIDI gericht.

Een XMF Meta-bestand bestaat uit verschillende nodes die hiërarchisch gegroepeerd zijn zoals een bestandssysteem met folders en bestanden. Bij XMF worden hiervoor respectievelijk de term containers en resources gebruikt. Een node kan ofwel een container ofwel een resource zijn. Een resource node bevat dan een verwijzing naar een intern bestand of een URL die verwijst naar een extern bestand.

3.3.4 MPEG-21Part 9 (File Format)

Binnen de MPEG-4 standaard (ISO/IEC 14496) (zie ook 3.6.5 MP4) zijn er verschillende delen die bestandsformaten definiëren voor de opslag van tijdsgebaseerde media, zoals audio en video. Deze zijn echter allemaal gebaseerd op en afgeleid van het ISO Basis Media File Formaat (ISO/IEC 14496-12), dat een hiërarchisch gestructureerde, media onafhankelijke definitie omvat die ook gepubliceerd is als deel van de JPEG2000-standaarden familie. Het bevat onder meer een basis containergestructureerd gedeelte en een definitie voor tijdssequenties van multimedia binnen een dergelijk containergestructureerd bestand.

Het MPEG-21 bestandsformaat (ISO/IEC 21000-9) gebruikt de structurele definitie van een containergebaseerd bestand, zoals het gedefinieerd is in het ISO Basis Media File Formaat, maar zonder de extra definities voor tijdsgebaseerde media. Het definieert de opslag van een MPEG-21 Digital Item (zie ook 5.5 MPEG-21/DIDL) plus alle eventueel bijkomstige (meta)data binnen datzelfde bestand, zoals foto's, filmpjes of andere niet-XML data. Containergebaseerde bestandsformaten laten immers toe om flexibele bestanden te creëren die meerdere containers bevatten die verschillende specificaties kunnen omvatten. Binnen MPEG-21 wordt een generieke meta-container gebruikt op bestandsniveau om enerzijds de beschrijving (MPEG-21 DID) van de resource weer te geven en verder ook een lijst met alle verwante resources, al dan niet ingebed in een sub-container of als extra verwijzing naar een compleet ander bestand. De volledige flexibiliteit en kracht van dergelijke URL-gebaseerde meta-container kan met volgend voorbeeld aangetoond worden:

- Items (andere bestanden) die nodig zijn voor het te beschrijven Digitale Item kunnen geïncludeerd worden binnen ditzelfde MPEG-21 bestand of in een ander bestand (al dan niet ook een MPEG-21 bestand).

- Items (bestanden) geïncludeerd in dit MPEG-21 bestand of in andere bestanden kunnen gefragmenteerd zijn en elk van die fragmenten kan daarbij ook interleaved zijn.
- Items (bestanden) kunnen beschermd zijn en er kan binnen het bestand aangegeven worden hoe daarmee om te gaan.
- Items (bestanden) kunnen een naam krijgen, waardoor er gemakkelijk kan naar verwezen worden binnenin een MPEG-21 bestand of zelfs vanuit een ander extern bestand.

3.3.5 OGM/OGG

OGM, of OGG Media, is een containerformaat dat een uitbreiding vormt op het OGG-containerformaat van Xiph.org. OGM voegt aan OGG onder andere de ondersteuning voor andere codecs toe dan diegene die ontworpen zijn door Xiph.org (Speex, Theora en Ogg Vorbis). OGM biedt namelijk ook ondersteuning aan videocodecs die gebruik maken van VfW en audiocodecs die ACM gebruiken. Net zoals bij videobitstromen wordt bij de audiobitstromen ondersteuning geboden voor een variabele bitrate. Algemeen wordt OGM als een tussenstap beschouwd tot de andere containerformaten, zoals Matroska, volgroeid zullen zijn en dezelfde mogelijkheden zullen bieden. Tot deze mogelijkheden behoren ondermeer ondersteuning voor hoofdstukken, meerdere ondertitels en meerdere audiokanalen.

3.3.6 Matroska (MKV/MKA)

Matroska is een open standaard multimediacontainerformaat dat gebaseerd is op EBML (Extensible Binary Meta Language). Dat is een binair bytegealinieerd formaat gebaseerd op de principes van XML.

Een Matroskabestand bestaat uit een header met informatie over de gebruikte EBML-versie en het bestandstype, in dit geval dus een Matroskabestand. De header wordt gevolgd door de Metaseek-sectie die de plaats aanduidt van de verschillende andere secties binnen het bestand. Dat is nodig omdat elke sectie in principe overal in het bestand kan voorkomen en men dus het hele bestand zou moeten parsen bij het zoeken naar informatie. Er zijn secties voorzien voor o.a. kanaalinformatie, hoofdstukinformatie en tags.

Matroska kent twee onderverdelingen: MKV, dat zowel video als audio kan bevatten, en MKA, dat enkel bedoeld is voor audio. Er is ondersteuning voor bijna alle video- en audioformaten zoals MPEG-1, MPEG-2, MPEG-4, Quicktime, Real, Theora voor video en MP1, MP2, MP3, PCM, AC3, FLAC, AAC voor audio. Hierbij worden zowel variabele audio bitrate als variabele framerate ondersteund. Verder maakt Matroska het ook mogelijk bestanden van om het even welk type toe te voegen. Zo kunnen bijvoorbeeld transcripties aan het bestand toegevoegd worden.

Matroska kan een onbeperkt aantal videostromen, audiostromen, afbeeldingen en ondertitels bevatten en laat ook toe lettertypes toe te voegen voor bijvoorbeeld de ondertitels. Matroska biedt verder ook een robuuste ondersteuning voor streaming, hoofdstukken en DVD-achtige menu's.

3.3.7 MXF

MXF, of Material eXchange Format, is een standaard containerformaat voor professionele video en audio en wordt gevormd door een set SMPTE-standaarden. MXF is een open bestandsformaat dat specifiek ontworpen werd om A/V-materiaal samen met de geassocieerde data en metadata tijdens de productiefase uit te wisselen. De ontwikkeling van MXF is een collaboratief proces van verschillende fabrikanten en organisaties Pro-MPEG, EBU en de AAF Association.

MXF is een veelzijdig bestandsformaat dat volgende taken aankan:

- Bewaren van eenvoudig afgewerkt materiaal en bijhorende metadata (tape replacement)
- Bewaren van materiaal in een streamable formaat dat het mogelijk maakt op het te bekijken terwijl het doorgestuurd wordt
- Verpakken en bewaren van een playlist met bestanden en de bijhorende synchronisatie informatie
- Verpakken van om het even welk compressieformaat
- Bewaren van cuts-only EDL's (Editing Decision List -een EDL bevat de gegevens zoals gebruikt bij audiovisuele content editingsystemen en doet dienst als een soort tijdslijn-) en het eigenlijke materiaal waarop het van toepassing is

Zowel real-time streaming (eindgebruikers kijken “live”) als bestandstransfers (tussen computersystemen onderling) zijn belangrijk in een veralgemeende A/V-constellatie. Daarom is het nodig dat beiden compatibel en interoperabel zijn. MXF is daarom ook zo ontworpen dat het ook een streaming formaat is, waardoor het naadloos een brug kan vormen tussen beide transfer types. MXF ondersteunt alle mogelijke video- en audioformaten en laat ook toe dat willekeurige bestanden worden toegevoegd. Dit laat toe transcripties, beelden, enz. toe te voegen.

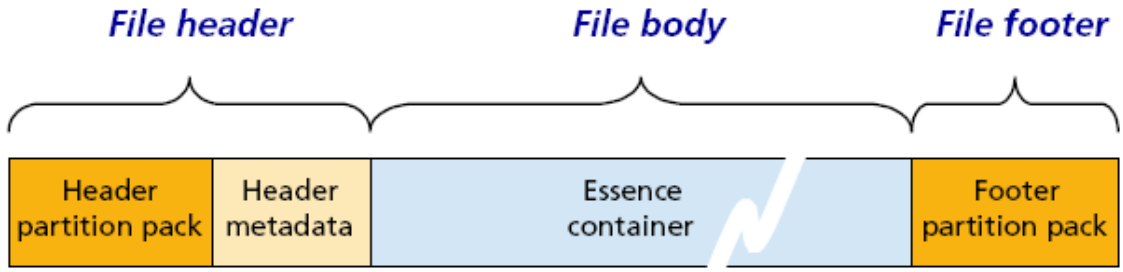
Zoals vermeld, was interoperabiliteit het belangrijkste objectief tijdens de ontwikkeling van MXF. Het is aldus:

- Cross-platform: het is volledig netwerkprotocol en besturingssyteem neutraal
- Compressie onafhankelijk: er worden geen converties uitgevoerd tussen verschillende codecs, maar het is gemakkelijk om verschillende codec-formaten, alsook ongecompresseerde data, in eenzelfde omgeving te beheren
- Een brug tussen streaming en transfers: volledig transparante, bidirectionele uitwisseling is mogelijk tussen al dan niet streamable bestanden

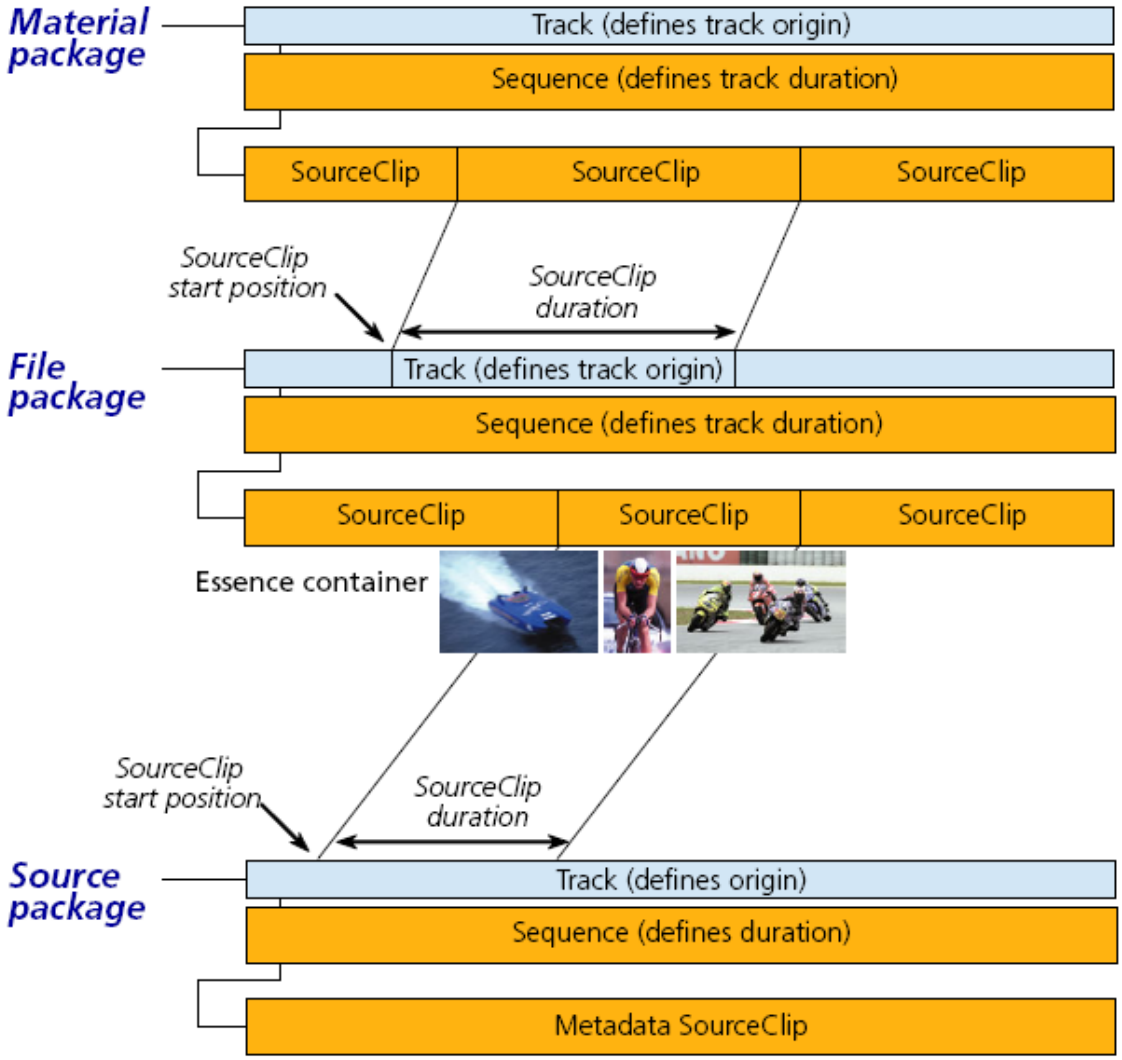
Een MXF-containerbestand bestaat uit een header, een footer en een body (met het eigenlijke A/V-materiaal). Elk item in een MXF-bestand is KLV-geencodeerd (Key Length Value) wat wil zeggen dat het uniek geïdentificeerd kan worden door een 16-byte sleutel en de lengte. Het kennen van de lengte van elk item binnen het MXF-bestand laat immers toe om eenvoudige decoders te implementeren en stukken “onbekende” data links te laten liggen tot er een (volgende) decoder gebruikt wordt die dat welbepaalde stuk data wel kan interpreteren. In het handige header metadata gedeelte van het MXF-bestand worden metadata, tijdsparameters en synchronisatie bewaard. Synchronistie informatie en de beschrijving van het materiaal wordt op drie verschillende niveaus bewaard:

- Material Package (MP): hier wordt de output tijdslijn van het bestand bewaard
- File Package (FP): het eigenlijke materiaal wordt hierin beschreven
- Source Package (SP): afgeleiden van dat materiaal (vb. EDL’s)

Elk van deze “packages” (MP, FP of SP) kan zijn eigen hoeveelheid tracks (audio, video en/of metadata) hebben. Elke track op zich kan een sequentie SourceClips bevatten.

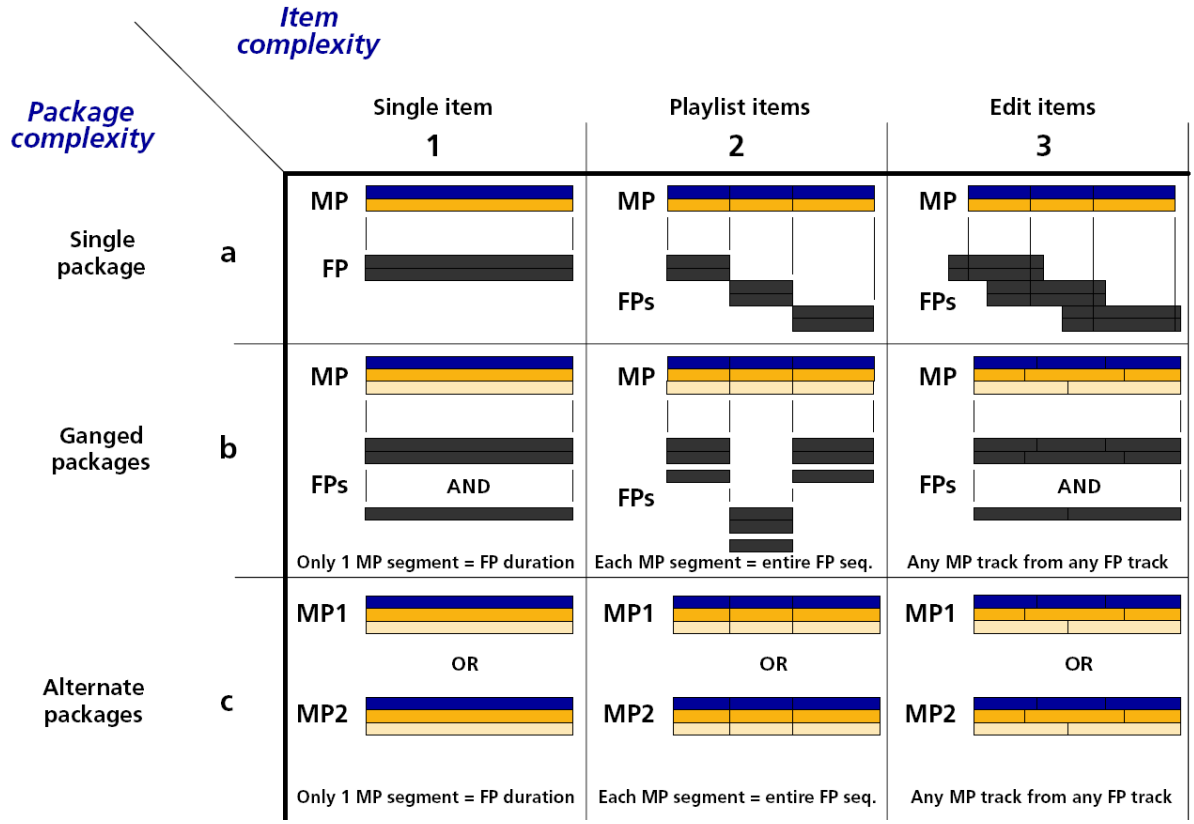


Simple MXF file structure



Om de complexiteit en het aantal vrijheidsgraden van MXF te kunnen beheren zijn er daarenboven ook een aantal “operationele patronen” (Operational Patterns) in het leven

geroepen. Hieronder zie je de grid die verticaal onderverdeeld wordt naargelang de complexiteit van de tijdslijn binnen het MXF-bestand en horizontaal onderverdeeld naargelang het aantal “packages” binnen datzelfde MXF-bestand.



MXF ondersteunt ook de toevoeging van metadata en enkele professionele functies zoals een volledige timecode platformonafhankelijkheid. MXF metadata kan volgende informatie bevatten:

- Bestandsstructuur
- Titel & keywords
- Ondertitels
- Referentienummers
- Editeernotities
- Versienummer
- Locatie, tijd en datum, …

3.3.8 MP4

MP4 of MPEG-4 part 14 is een multimedia containerformaat dat een onderdeel vormt van de MPEG-4-standaard. MP4 kan zowel audio- als videostromen bevatten. Hierbij ondersteunt MP4 de standaard videoformaten MPEG-1, MPEG-2, MPEG-4 en MPEG-4 AVC. Voor audio worden de standaarden (HE)-AAC, MP3, MP2, MP1, CELP, TwinVQ, Vorbis en Apple Lossless ondersteund. Wanneer een MP4-container enkel audio bevat, krijgt deze vaak de extensie M4A. Deze extensie wordt onder andere voor podcasts gebruikt.

Naast de gewone audio- en videostromen kent MP4 ook zogenaamde private stromen. Deze private stromen kunnen om het even welke gegevens bevatten. Zo gebruikt Nero deze stromen om ondertitels in Dvd-formaat toe te voegen.

MP4 ondersteunt verder ook afbeeldingen, hyperlinks, ondertitels, hoofdstukken, variabele audio bitrate en variabele framerate.

3.3.9 3GP

3GP, of 3G Protocol, is een multimedia containerformaat, ontworpen door de Third Generation Partnership Project (3GPP), voor gebruik met 3G mobiele telefoons. 3GP is een vereenvoudigde versie van het MP4-containerformaat en is ontworpen met als doel de vermindering van opslag- en bandbreedtevereisten. 3GP ondersteunt zowel MPEG-4 Part 2, H.264/AVC als H.263 voor video en AMR-NB, AMR-WB, AMR-WB+ en (HE)-AAC-LC voor audio. 3GP biedt ook ondersteuning aan variabele audio bitrates, variabele framerates en ondertitels. 3GP bestanden kunnen zowel gestreamd als gedownload worden (zie bijvoorbeeld MMS-berichten).

3.3.10 ASF

ASF, of Advanced Systems Format, is een propriëtair containerformaat ontworpen door Microsoft als onderdeel van het Windows Media Framework. De vroegere naam Advanced Streaming Format geeft het hoofddoel weer van het containerformaat: streaming. ASF kent twee versies. Versie 1.0 is veruit de meest gebruikte versie maar is gesloten, de opbouw is dus op enkele details na niet gekend. Versie 2.0 is open maar wordt amper gebruikt.

ASF ondersteunt bijna alle video- en audioformaten die werken via VfW en ACM maar wordt meestal gebruikt in combinatie met Microsofts eigen formaten. Verder ondersteunt ASF ook metadata zoals artiest en titel, variabele audio bitrate, variabele framerate, hoofdstukken en

ondertitels. ASF biedt ook foutcorrigerende technieken en een digital rights management framework.

3.3.11 MOV

MOV is een multimedia containerformaat ontworpen door Apple als basis voor het MP4-containerformaat. Deze container kan zowel video, audio als hoofdstukken bevatten en ondersteunt variabele audio bitrate en variabele framerate. MOV ondersteunt alle formaten die de Quicktime codecmanager ondersteunt, zoals MPEG-4 en de Sorensen codec, en alle audioformaten die de soundmanager en coreaudio ondersteunen, zoals AIFF, WAV en MP3.

In een MOV-container kan elk kanaal voorgesteld worden door de mediastroom zelf of door een referentie naar de mediastroom in een ander bestand. Binnen de MOV-container worden de kanalen in een hiërarchische structuur van atomen geplaatst. Deze atomen kunnen ofwel “ouder” zijn dan andere atomen ofwel zelf media of data bevatten.

MOV-containers bevatten een tijdslijn die los staat van de mediastromen. Hierdoor kunnen MOV-containers eenvoudig worden aangepast zonder dat de mediastromen moeten worden gekopieerd.

3.3.12 AVI

AVI, of Audio-Video Interleaved, is een multimedia containerformaat dat ontworpen is door Microsoft [6.1.2.8-1]. AVI-containers kunnen meerdere audio- en videokanalen bevatten. Een AVI-container bestaat uit een header met informatie over de video, zoals breedte, hoogte en framerate, en de eigenlijke data. Verder kan een container ook een index bevatten die toelaat te navigeren binnen de container. AVI-containers ondersteunen bijna alle audio- en videoformaten die beschikbaar zijn via DMO, ACM en VfW. AVI ondersteunt variabele audio bitrates maar met enkele beperkingen (niet via ACM) en variabele framerates. Ondertitels en hoofdstukken worden ook ondersteund via modificaties, maar dan buiten Microsoft.

3.3.13 FLV

FLV, of Flash Video, is een propriëtair containerformaat ontworpen door Adobe dat onder meer door Google Video en YouTube gebruikt wordt. FLV kan slechts één video en één audiostroom bevatten per bestand. Verder kan een container ook Flashcontent bevatten. FLV ondersteunt de videoformaten Sorensen, VP6 en Screen Video en de videoformaten

MP3, Nellymoser, ADPCM en PCM. Een FLV-container kan op verschillende manieren bij de eindgebruiker terechtkomen: via download, embedded in een flash animatie of door streaming via het RTMP-protocol.

In de nieuwe versie van FLV wordt ook ondersteuning aan H.264/AVC en HE-AAC geboden.

3.3.14 RealMedia

RealMedia is een multimedia containerformaat ontworpen door RealNetworks. Realmedia is een populair formaat voor het streamen van audio en video via het internet. RealMedia ondersteunt de videoformaten RealVideo 8-9-10 en de audioformaten HE-AAC, Cook, Vorbis en RealAudio Lossless. Verder biedt Realmedia ook ondersteuning aan variabele framerate, ondertitels en met behulp van de RMVB-extensie ook aan variabele bitrates.

4 Informatie over de data

4.1 Inleiding

Voor de term ‘metadata’ bestaan verschillende definities maar doorgaans worden metadata omschreven als “bits about bits”: “data over data”. Metadata bieden (gestructureerde) informatie over een bron (resource). Onder bronnen verstaan we alle mogelijke objecten of subjecten waarover informatie kan worden opgeslagen. Voorbeelden zijn onder meer tekst, fysieke objecten, software maar ook personen, gebeurtenissen of diensten.

Naargelang de soort informatie die de metadata bevatten, kan men verschillende types onderscheiden: administratieve metadata (rechten, plaats,...), beschrijvende metadata en bewaringsinformatie (toestand, verhuizingen,...), technische informatie (formaat, encryptie,...) en gebruik.

Een meer geavanceerde vorm van metadata geeft ook onderlinge relaties aan, bijvoorbeeld de creatie of publicatie. In dit voorbeeld worden de persoon “Interviewer” en de bron “mondelinge historische bron” aan elkaar gelinkt door de relatie “creatie”.

Bovendien kunnen metadatastandaarden worden ingedeeld volgens zoekmogelijkheden. Naast gewone metadatastandaarden (zoals MARC/MARC21) bestaan er namelijk ook semantische standaarden die ‘intelligente’ zoekmethoden ondersteunen en daarbij rekening houden met de betekenis van zoektermen of gebruik maken van thesauri. Een bekend probleem bestaat er bijvoorbeeld in dat “computers” gegevens zoals “H. Claus” en “Hugo Claus” als twee verschillende personen beschouwen. Om dergelijke beperkingen te ondervangen, worden vaak “woordenboeken” gebruikt met afgesproken termen. Ook thesauri kunnen dus helpen door bij zoekacties termen met een gelijkaardige betekenis toe te voegen aan de request.

Metadata kent verschillende functies. De belangrijkste is dat metadata het terugvinden van relevante informatie vergemakkelijkt. Het kan helpen om electronische bronnen te organiseren en de interoperabiliteit te verzekeren, een digitale identificatie te voorzien en het archiveren en preserveren te ondersteunen.

Er bestaan op dit moment ontzettend veel metadatastandaarden en de keuze van de te gebruiken metadatastandaard is lang niet gemakkelijk. De metadatastandaarden verschillen in granulariteit bij het beschrijven, semantiek en toepassingsgebied.

De meest gebruikte en waarschijnlijk ook de meest eenvoudige metadatastandaard is Dublin Core. Het is zowat de lingua franca wat betreft metadatastandaarden. De kracht van deze

standaard is zijn simpliciteit en generaliteit. De standaard bestaat uit 15 velden. Met deze velden kan elke bron beschreven worden, maar deze beschrijving is vaak te beperkt. Daarom wordt Dublin Core veel gebruikt als een additionele metadatastandaard naast een andere metadatastandaard die de bronnen veel nauwkeuriger beschrijft. Omdat de meeste systemen overweg kunnen met Dublin Core, zorgt een mapping van je metadatastandaard naar Dublin Core voor de nodige interoperabiliteit. Doordat de 15 velden van Dublin Core optioneel en herhaalbaar zijn, kan zowat elke metadatastandaard gemapt worden naar Dublin Core, eventueel wel met verlies van data tot gevolg aangezien niet alle velden kunnen gemapt worden naar de 15 velden van Dublin Core.

Nu volgt een overzicht van de veel gebruikte metadatastandaarden in het bibliotheekwezen, de omroepsector, de culturele sector en de archiefsector. Deze indeling is echter niet arbitrair. De sectoren vertonen nogal veel overlap en bepaalde metadatastandaarden zijn toepasbaar in de verschillende sectoren. Daarom werd er voor gekozen om in de inleiding van dit hoofdstuk een overzicht te geven van de meest gebruikte standaarden per sector, terwijl verder in het hoofdstuk een uitvoerigere bespreking wordt geven van de metadatstandaarden. Deze besprekingen worden ondersteund met een voorbeeld van een metadatarecord. Deze record is beschreven in XML of RDF\XML, die overweg kan met semantiek.

Wat betreft de omroepsector zijn er op dit moment twee metadatastandaarden die veel gebruikt worden: P/META en SMEF-DM. P/META is de standaard die in Vlaanderen binnen de omroepsector het meest wordt gebruikt. P/META is een metadatastandaard die de nodige velden heeft om de uitwisseling van informatie die typisch gekoppeld is aan audio-visueel materiaal te voorzien. In feite wordt zowel binnen de commerciële als publieke omroepsector gebruik gemaakt van het IPEA-model dat een subset is van P/META voor de uitwisseling van programmagegevens. Het IPEA-model is ook ontwikkeld in samenwerking met de Vlaamse omroepen. SMEF-DM is een gelijkaardige inspanning geleverd door de BBC.

Binnen het bibliotheekwezen worden MARC/MARC21 en FRBR het meest gebruikt. MARC is een standaard voor de representatie en de communicatie van bibliografische en aanverwante informatie. De hoofdfunctie van de standaard was dan ook de vereenvoudiging en bespoediging van het terugvinden van boeken in de bibliotheek. Het MARC-formaat biedt een hoge graad aan granulariteit, wat de standaard ook complex maakt. Deze granulariteit zorgt ervoor dat de bron zeer nauwkeurig kan worden beschreven. Het formaat is desondanks compact, doordat de veldnamen zoals b.v. “plaats van publicatie” immers worden vervangen door een korte code, wat de leesbaarheid van de standaard echter

bemoeilijkt. FRBR is een conceptueel model voor gebruik binnen de bibliografische wereld waarbij de nadruk meer op de eindgebruiker ligt. Het model is ontwikkeld om bepaalde gebruikersactiviteiten te vergemakkelijken zoals b.v. het terugvinden van records. De bibliografische entiteiten die worden gedefinieerd binnen dit model worden onderverdeeld in groepen, die op hun beurt verder kunnen worden onderverdeeld. Deze standaard is dus ook zeer granulair en laat dus ook een nauwkeuruge beschrijving toe van de entiteiten. Er moet echter opgemerkt worden dat het gebruik van een zeer granulaire metadatastandaard een zekere implementatiekost met zich meebrengt.

CDWA beschrijft de data uit kunstdatabanken aan de hand van een conceptueel raamwerk voor het beschrijven en opvragen van informatie over kunstwerken, achitectuur of ander cultureel materiaal. Deze standaard wordt dan ook voornamelijk in de cultuursector gehanteerd. Het CDWA bevat 512 categorieën en subcategorieën. Een kleine subset van deze categorieën vormt de core. Deze categorieën stellen de minimale informatie voor die nodig is om een werk te beschrijven en te identificeren. Van deze core bestaat er een XML-schema, CDWA Lite genoemd, die bijdraagt tot de implementatie van dit schema. Dit model is ook in overeenstemming met OAI-PMH die het uitwisselen van gegevens tussen verschillende bibliotheken vergemakkelijkt. CIDOC-CRM is een andere standaard die gangbaar is in de cultuursector. Het CIDOC Conceptueel Referentie Model (CRM) levert de definities en een formele structuur om concepten en relaties die gebruikt worden bij de documentatie van cultureel erfgoed te beschrijven. CIDOC CRM richt zich ook voornamelijk op de beschrijving van de contextuele informatie. Dit houdt voornamelijk de historische, geografische en theoretische achtergrond in van de tentoongestelde items, waardoor hun waarde en betekenis vergroten. Deze standaard wordt ook vaak gebruikt als metadataspil, zoals dat het geval is bij Dublin Core, om de interoperabiliteit te vergroten van het systeem. Waar Dublin Core dit doet met veel gegevensverlies, is CIDOC-CRM uitgebreid genoeg om de interoperabiliteit aan te leveren met een minimum aan gegevensverlies.

Binnen de archiefsector is ISAD(G) de standaard. Deze standaard moet helpen bij het opstellen van beschrijvingen van collecties en objecten. De standaard bestaat uit verschillende regels, maar voorziet niet in een eigen encodering en is dus eerder een “handleiding” voor het beschrijven van collecties. De regels houden b.v. richtlijnen in voor multi-level beschrijvingen, het gebruik van referenties, titels, dateringen en dergelijke.

4.2 Descriptieve Metadatastandaarden

4.2.1 Dublin Core

Dublin Core is een sectoroverschrijdende metadatastandaard. De standaard is niet ontworpen met het doel een verfijning en complexiteit van bijvoorbeeld het MARC-formaat te evenaren. Bij Dublin Core wordt namelijk getracht een grootste gemene deler te vormen tussen metadatastandaarden die in verschillende sectoren worden toegepast, met het oog op de vereenvoudiging van onderlinge informatie-uitwisseling en zoekopdrachten. Bij de invulling van metadata moet men er daarom rekening mee houden dat de semantiek van elementen verschilt naargelang de sector.

In Dublin Core wordt gesproken over *Resources*, *Elementen*, *Qualifiers* en *Schemes*. *Resources* zijn de objecten die beschreven worden met behulp van 15 *elementen*, waaronder *creator* en *rights*. Zo geeft het element *type* de aard van het object weer, bijvoorbeeld *sound* waartoe ook mondelinge historische bronnen behoren. Deze 15 elementen vormen het zogenaamde Dublin Core Simple-profiel.

Het Dublin Core Qualified voegt aan Dublin Core Simple drie extra elementen toe (waaronder het doelpubliek) en vult het profiel ook verder aan met *Qualifiers* en *Schemes*. *Qualifiers* worden gebruikt om elementen te verfijnen (bijvoorbeeld dat het element *creator* een fotograaf of een auteur aangeeft). Deze *Qualifiers* zijn niet aan regels gebonden. Dit heeft tot gevolg dat niet elke software deze facultatieve qualifiers zal begrijpen. Software die niet de term “fotograaf” niet herkent, interpreteert die dan als een *creator*. *Qualifiers* bezorgen software die ermee overweg kan, extra informatie zonder aan compatibiliteit in te boeten.

De toevoeging *Scheme* laat vervolgens toe aan te geven hoe elementen moeten worden ingevuld. Zo kan worden aangegeven dat het *subject* een trefwoord betreft uit een bepaalde thesaurus en niet een vrij sleutelwoord. In het attribuut *Scheme* wordt dan de gevolgde thesaurus vermeld.

Behalve de standaardset van elementen kunnen ook andere elementen worden toegevoegd. Het is wel aan te raden elementen te gebruiken die afkomstig zijn van andere metadatastandaarden. Een nieuwe set elementen vormt dan een *application profile*.

Een voorbeeld van een Dublin Core Entry:

```
<rdf:RDF
   xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
```

```
   xmlns:dc="http://purl.org/dc/elements/1.1/">
   <rdf:Description rdf:about="http://media.example.com/audio/guide.ra">
      <dc:creator>X</dc:creator>
      <dc:title>Interview met een oudstrijder</dc:title>
        <dc:description>Interview met een oudstrijder 1940-
        1944</dc:description>
      <dc:date>1999</dc:date>
   </rdf:Description>
</rdf:RDF>
```

Nadelen van Dublin Core:

- Dublin Core beperkt zich tot de beschrijving van resources zoals boeken en geluidsfragmenten maar ondersteunt niet de beschrijving van personen en instellingen.
- Dublin Core beschrijft voornamelijk het voorwerp zelf maar in beperkte mate het uitgebeelde/beschreven onderwerp.
- Verschillende interpretaties van eenzelfde element kunnen leiden tot “vertaalproblemen”, hoewel dit in feite alle metadatastandaarden typeert.
-

Voordelen Dublin Core:

- Dublin Core vereenvoudigt aanzienlijk het samenvoegen van metadata van instellingen die gebruik maken van deze standaard.
- Het nadeel van het beperkt aantal elementen van Simple Dublin Core kan weggewerkt worden door het gebruik van *qualifiers*.
- De standaard ondersteunt RDF-gebaseerde opslag.

4.2.2 MPEG-7

MPEG-7 werd ontworpen door de Motion Pictures Expert Group (MPEG). Deze werkgroep is vooral bekend voor zijn standaarden voor de codering van video en audio. MPEG-7 focust echter op de representatie van informatie over de content in plaats van op de content zelf. MPEG-7 wil een rijke verzameling aan gestandaardiseerde hulpmiddelen bieden voor het beschrijven van multimediale content. De beschrijving van content moet mogelijk zijn ongeacht de wijze van opslag, de codering, de technologie, enz. Zo kan een beschrijving zowel handelen over een geprinte foto als over een interview in een digitaal audioformaat.

MPEG-7 bestaat uit Descriptors, Multimedia Description Schema's, Description Definition Language en hulpmiddelen die de binarisatie, de synchronisatie, het transport en de opslag van de descriptors voor hun rekening nemen.

Een Descriptor is de voorstelling van een kenmerk. Deze voorstelling is zowel syntactisch als semantisch vastgelegd. Een object heeft natuurlijk meerdere kenmerken en een uniek object kan uiteraard door meerdere descriptors worden beschreven. Multimedia Description Schema's zorgen voor de weergave van de structuur en de semantiek van de relaties tussen de verschillende descriptors, maar ook tussen andere Description Schema's.

Voor het definiëren van de structurele relaties tussen de descripties wordt gebruik gemaakt van een XML-gebaseerde taal, de Description Definition Language. Hiermee kunnen beschrijvende schema's gecreëerd en aangepast worden.

MPEG-7-beschrijvingen laten toe verschillende dieptes van detail weer te geven. Hierdoor is het mogelijk bepaalde informatie weg te laten of verder te verfijnen. Deze verfijningen verschillen natuurlijk per toepassingsgebied. Voor historische audiobronnen zou een beschrijving op hoog niveau kunnen zijn: "Interview met een oudstrijder". Op lagere niveaus kan dan gedetailleerdere informatie worden meegegeven. Zo kan de beschrijving op lager niveau uitgebreid worden met de naam van de oudstrijder, informatie over de oorlog, enz.

Naast de beschrijvingen over de inhoud van een object kan ook extra informatie toegevoegd worden:

- Informatie over de creatie- en productieprocessen van de content
- Informatie over het gebruik van de content zoals copyright informatie en raadplegingen in het verleden
- Informatie over het opslagformaat van de content
- Informatie over collecties, interactie van de gebruiker met de content, enz.

Voordelen:

- MPEG-standaard

Nadelen:

- Voorlopig weinig industriële interesse vanwege te ingewikkeld -1182 elementen, 417 attributen en 377 complexe types-, alhoewel de verschillende "afgeslankte" profielen daarvoor een uitkomst probeerden bieden

- Voorlopig weinig industriële interesse vanwege te flexibel, waardoor het weinig interoperable wordt. Zo is het mogelijk eenzelfde modulaire beschrijvingen te geven binnen verschillende abstractieniveaus, descriptors kunnen aan een arbitrair segment toegevoegd worden met om het even welk detailniveau en het huidige schema kan zelfs onbeperkt verder uitgebreid worden
- Nog steeds worden wijzigingen voorgesteld, specifiek voor het Query Format

MPEG-7 voorbeeld:

```
<Mpeg7>
    <Description xsi:type="CreationDescriptionType">
        <CreationInformation id="track4">
            <Creation>
                <Title type="songTitle">Interview met oudstrijder</Title>
                <Abstract>
                    <FreeTextAnnotation>Interview over het leven van een
                    oudstrijder</FreeTextAnnotation>
                </Abstract>
                <Creator>
                    <Agent xsi:type="PersonType">
                         <Name>
                                 <FamilyName>De Smedt</FamilyName>
                                 <GivenName>Jan</GivenName>
                         </Name>
                    </Agent>
                    <CreationCoordinates>
                         <Date><TimePoint>1999</TimePoint></Date>
                    </CreationCoordinates>
                </Creator>
            </Creation>
        </CreationInformation>
    </Description>
</Mpeg7>
```

4.2.3 P/META:

Toepassingsgebied en opzet:

Bij de uitwisseling van de inhoud van programma's heeft het gebruik van metadata altijd al een cruciale rol gespeeld. Omdat de drang naar interoperabiliteit bij moderne systemen steeds groter wordt, groeit ook de nood aan projecten die een toenemende standaardisatie van die metadata bevorderen. Het EBU P/Meta project, opgestart in 1999, is een van de

belangrijkste actoren actief op dit gebied met als bijzonderheid dat er bij het zoeken naar een metadatastandaard vanuit het standpunt van de omroepen vertrokken wordt.

EBU heeft als doel het ontwikkelen van een standaard metadata uitwisselingsraamwerk dat kan ingezet worden voor het delen van de betekenis van elektronische informatie die noodzakelijk of nuttig is voor de business-to-business (B2B) uitwisseling van programmagerelateerde informatie en eigenlijke inhoud. En dit zonder dat de interne structuur, de werkmethodiek of de algemene concepten van de participerende organisaties gewijzigd moeten worden. Organisatiespecifieke opslagschema's voor data kunnen gefilterd en afgebeeld worden op het P/Meta Schema zodat de uitwisseling van metadata tussen verschillende organisaties mogelijk gemaakt wordt, onafhankelijk van de onderliggende technologische infrastructuur voor datatransport.

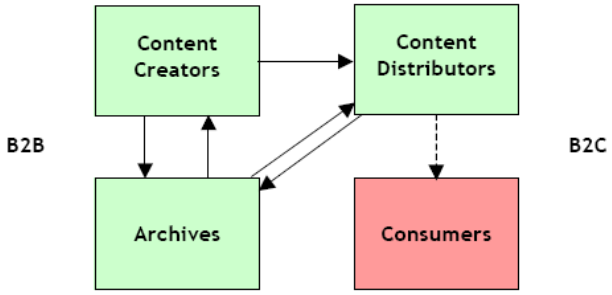
Wat is P_META?

Het P/Meta Schema is een verzameling van definities die voorziet in een semantisch raamwerk voor de uitwisseling van informatie die typisch gekoppeld is aan audiovisueel materiaal. Het schema bevat de identificatie van concepten en subjecten tijdens een data-analyse fase, waarnaar dan gerefereerd kan worden door middel van P/Meta Identifiers en Names. Ze worden geïdentificeerd met het oog op het nastreven van een maximale nauwkeurigheid bij het opmaken van beschrijvingen, een maximale flexibiliteit in het gebruik en hergebruik, en een maximale flexibiliteit bij de definitie van basiselementen en datastructuren.

Wanneer we over de uitwisseling van metadata spreken dan hebben we het eigenlijk over een proces dat actief is op drie niveaus: de Definition Layer, de Technology Layer en de Data Interchange Layer. De eerste laag, ook als Descriptive Metadata Layer of de Semantic Layer aangeduid, betreft de semantiek van de informatie-elementen en is de laag die door P_META wordt ingevuld. Concreet bedoelen we hiermee de exacte definitie en betekenis van elk beschrijvend element dat van belang is bij de meeste productieprocessen. Deze definities berusten op de professionele betekenis en interpretatie van elk concept en worden uitgedrukt in normale menselijke taal. De invulling van de Technology Layer hangt af van de gebruikte technologie voor de uitwisseling van informatie. Om het even welke gekozen technologie moet bij de uitwisseling van informatie de originele betekenis van de gedefinieerde informatie-elementen behouden. Voorbeelden van technologieën zijn bijvoorbeeld XML, KLV of simpele tekstdocumenten. De Data Interchange Layer behelst hoe en via welk medium de gecodeerde informatie effectief uitgewisseld wordt.

Context

P/Meta is hoofdzakelijk van toepassing in een B2B omgeving. Ondanks de focus op B2B wordt de interoperabiliteit met business-to-consumer (B2C) metadata beschouwd als een elementaire eigenschap van P/Meta. Onderstaande figuur illustreert het gangbare procesmodel dat drie verschillende business actoren en een groep Consumers identificeert, met alle mogelijke onderlinge interfaces voor de uitwisseling van informatie.



Content Creators, ook wel Producers, houden zich bezig met de productie van programma's en andere media en zorgen ervoor dat nieuw materiaal beschikbaar is voor publicatie. De entiteit Archives omvat het zorgvuldig bewaren en beschermen van bestaand materiaal. Het maakt het integraal hergebruik van materiaal mogelijk en kan worden aangewend als bijkomende bron voor de creatie van nieuwe programma's. Content Distributors faciliteren de publicatie en de levering van materiaal aan de eindgebruikers. Zij nemen deel aan uitwisselingen in het B2C scenario. Content Distributors worden ook wel als Broadcasters of Content Aggregators aangeduid. De Consumers zijn de gebruikers aan het eind van de media distributieketen.

Objectieven

Wat het B2B scenario betreft, worden vier doelstellingen omlijnd:

- de identificatie en de erkenning van materiaal mogelijk maken

- het verstrekken van een beschrijving van materiaal geschikt voor gebruik op de redactie

- het vastleggen van de rechten met betrekking tot het materiaal

- het aanbieden van een minimale hoeveelheid technische details om de eigenlijke uitwisseling en het correcte gebruik van materiaal mogelijk te maken

Het P/Meta model

De P/Meta standaard introduceert een gelaagd hiërarchisch model bestaande uit vijf exchange concepts: een Programme Group, een Programme, een Items Group, een Item of Programme Item en een Media Object of MOB. De EBU definieert de vijf concepten als volgt:

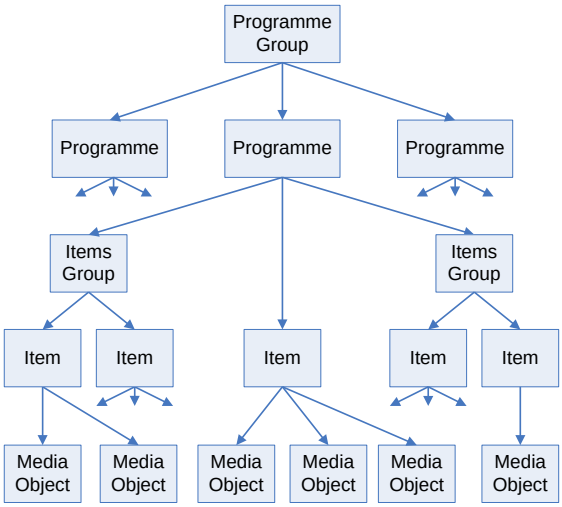
- Programme Group: een verzameling van programma's gecreëerd door een beslissing van een commissie of opgelegd door het volgen van een vooropgestelde planning, gebonden door een gemeenschappelijk redactioneel concept, e.g. VRT Nieuws.

- Programme: een audiovisueel werk dat gedefinieerd en gecreëerd werd door de beslissing van een commissie, e.g. Het Journaal.

- Items Group: Items kunnen gegroepeerd worden in een Items Group, een verzameling bestaande uit Programme Items dat een samenstellend onderdeel van een programma vormt of gebonden is door een redactiebeslissing. [9] Wanneer een Item in eenzelfde nieuwsuitzending meer dan eens aan bod komt, dan worden beide Items gebundeld in een Items Group, e.g. aanvankelijk toont men een vooraf opgenomen reportage, enkele Items verder heeft men een live interview rond hetzelfde onderwerp.

- Item of Programme Item: een samenstellend onderdeel van een programma, zo gedefinieerd door een besluit van een redactie. Het kan op zichzelf geïdentificeerd worden of door zijn positie in een programma. Het is bijvoorbeeld afgebakend in tijd, e.g. een item uit een nieuwsuitzending.

- Media Object of MOB: één enkele component van één enkel media type van een Programma of een Item. Het is continu in de tijd en fysiek van aard in tegenstelling tot bovenstaande logische concepten, e.g. een bestand dat de daadwerkelijke video van een nieuwsuitzending bevat.



Andere entiteiten worden in de Definition Layer door hun context bepaald. Een van de belangrijkste entiteiten is een Brand, een collectie van assets met een herkenbare collectieve identiteit, e.g. één of canvas. Twee andere belangrijke entiteiten zijn personen en organisaties betrokken bij de creatie, het beheer en de controle van inhoud.

Lijst van P/Meta componenten

Het vastleggen van de semantiek van metadata voor de uitwisseling van audiovisueel materiaal wordt binnen de P/Meta standaard bereikt door het opstellen van een lijst van Attributes, een lijst met referentiedata en een lijst Transaction Sets.

Binnen P/Meta is een attribuut het meest eenvoudige element dat informatie kan bevatten. Het is mogelijk om een attribuut uniek te identificeren door gebruik te maken van een code en een naam, e.g. het attribuut met code ‘A1’ en naam ‘ADDRESS_DELIVERY_CODE’. Door de standaard wordt de courante betekenis van elk attribuut, het type van de waarde (e.g. Boolean, Integer, Uncontrolled Text, Controlled Code, …), een externe referentie, gekende aliassen en in sommige gevallen enkele voorbeelden, op ondubbelzinnige wijze vastgelegd. De betekenis van een attribuut kan worden verrijkt door de context waarin het wordt gebruikt. Hieruit volgt dat hetzelfde attribuut in verschillende contexten kan worden (her)gebruikt en dit om uiteenlopende B2B doelen te dienen. Als voorbeeld beschouwen we het attribuut Language Code: het kan worden gebruikt om de taal van de originele dialoog van een Item aan te geven en ook om de taal van de eigendomsrechten aan te duiden.

Voor elk attribuut dat als waardetype een Controlled Code heeft, moet een lijst voorhanden zijn die alle toegestane waarden bevat, samen met de exacte betekenis van elke waarde.

Dergelijke waardelijsten kunnen direct vanuit P/Meta aangeboden worden of afkomstig zijn van een externe bron als EBU, ISO of SMPTE.

Naast de attributen, werkt P/Meta ook met Transaction Sets of functionele groeperingen van P/Meta attributen en/of andere P/Meta Sets. Een voorbeeld is de set 'S12 PERSON_DETAILS' die bestaat uit enkele attributen, zoals 'A89 PERSON_LAST_NAME' en 'A88 PERSON_FIRST_NAME', aangevuld met de subset 'S13 ADDRESS'. De P_META standaard bevat een aantal vooraf gedefinieerde Sets die geconstrueerd zijn als bouwstenen voor het opzetten van gemeenschappelijke data-uitwisselingen. Ook kunnen nieuwe logische sets gecreëerd worden om aan de eigen specifieke transactievereisten te voldoen. Deze sets worden geconstrueerd volgens een syntax en notatie aangegeven door de standaard. De voorgedefinieerde Sets zijn ontwikkeld om in de volgende domeinen alvast een generieke oplossing te bieden: metadata voor identificatie en erkenning, beschrijvende metadata, technische metadata, metadata met betrekking tot transacties, transmissies, rechten en andere.

Het P/Meta schema is een schema dat toelaat de logische inhoud en de betekenis van informatie te beschrijven los van de technische implementatie. Er kan aan de eisen van het P/Meta schema voldaan worden onafhankelijk van de keuze voor een bepaald platform, een coderingsstandaard, een transactieprotocol of zelfs, wanneer we met Controlled Code attributen te maken hebben, onafhankelijk van de gekozen taal.

IPEA: Innovatief Platform voor Elektronische Archivering

Twee grote Vlaamse omroepen, de commerciële omroep VMMa en de publieke omroep VRT, het faciliterende bedrijf Videohouse en verschillende universitaire onderzoeksgroepen geassocieerd met het IBBT, bundelden hun krachten in het IPEA project. Aanleiding voor het project was de nood aan informatie met betrekking tot het overstappen van een tapegebaseerde productie –en archiveringsomgeving naar een compleet bestandsgebaseerde workflow. Het IPEA project, lopende van januari 2005 tot december 2006, had o.a. tot doel het ontwikkelen van een algemene, gedeelde standaard voor de uitwisseling en de archivering van audiovisuele data.

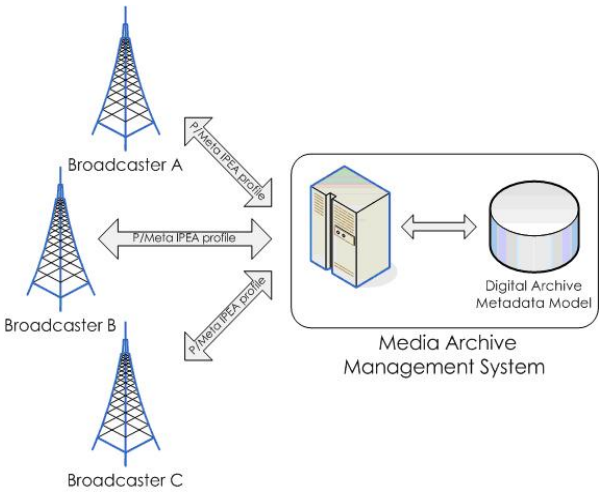
Een van de cruciale aspecten van het IPEA project was het creëren van een gemeenschappelijk metadatamodel, gerealiseerd binnen werkpakket vier (WP4: Creation of a standardized metadata model). Onder het metadatamodel werden de volgende zaken verondersteld:

- een gestandaardiseerde semantische beschrijving

- een gestandaardiseerde syntax

- de definitie van nuttige ontologieën

WP4 leidde tot de definitie van twee metadatastandaarden: een intern model en een uitwisselingsmodel. De eerste standaard, het intern model, is een ERD dat de layout en de relaties van het digitale archief definieert. Het tweede wordt gebruikt door externe gebruikers voor de invoer van media in een digitaal media-archief alsook voor het ophalen van media uit een dergelijk archief. In dit model wordt gebruik gemaakt van het IPEA profiel, i.e. een uitwisselingsstandaard die bestaat uit een subset van de EBU P/Meta 1.1 standaard. De subset werd gekozen om te voldoen aan de noden van de eerder vermelde Vlaamse omroepen. Opteren voor een internationaal genormeerde specificatie als P/meta vereenvoudigde het definiëren van de semantiek, de syntaxis en de schrijfwijze van de taal waarmee tussen verschillende actoren over programma's gecommuniceerd kan worden.



Door gebruik te maken van een subset van de P/Meta standaard kan men de semantiek en de syntaxis van alle elementen voor de interface strikt vastleggen. P/Meta zorgt voor een eenduidige definitie van de semantiek door het aanbieden van een exhaustieve lijst van alle attributen die belangrijk kunnen zijn om een programma te beschrijven en het specificeren van de mogelijke waarden voor alle discrete attributen. P/Meta bepaalt ook de syntaxis van de interface doordat de standaard aangeeft hoe met de voorgenoemde attributen zinvol gecommuniceerd kan worden. De syntaxis is zo opgesteld dat ze gemakkelijk kan worden geïnterpreteerd door machines en is daarom nogal abstract.

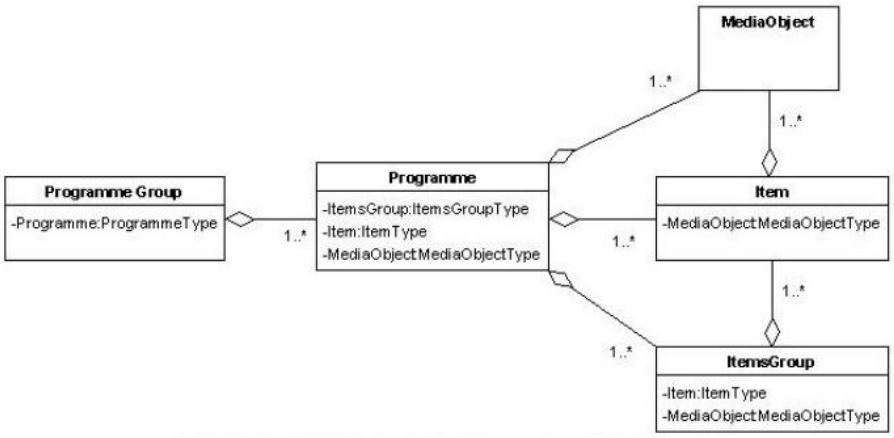
Echter, doordat P/Meta bewust abstractie maakt van de implementatie van de standaard en dit om onafhankelijk te blijven van de gebruikte technologie, moest hieromtrent binnen het IPEA project wel een afspraak worden gemaakt. Men koos ervoor om het IPEA profiel te implementeren door middel van XML en de gecreëerde documenten te controleren door XML schema's.

In november 2005 lanceerde de European Broadcasting Union P/Meta versie 1.2, die, mede onder impuls van de resultaten van het IPEA project, een paar Attributes en Sets aan versie 1.1 toevoegde.

P_META 2.0

Sinds juli 2007 is versie 1.2 vervangen door P/Meta 2.0. Deze versie is het resultaat van een aantal significante wijzigingen. De meest opvallende zijn de toewijzing van namen aan structuren die eerder naamloos waren, de optimalisering van de datastructuur voor het gebruik van XML en de opsplitsing van de originele P/Meta metadataset in een algemene toolkit en een aantal afzonderlijke toepassingsgerichte specificaties, e.g. voor de uitwisseling van programma's. Alle aanpassingen zorgen ervoor dat versie 2.0 niet meer direct compatibel is met de eerdere versies.

In P_META 2.0 wordt een Brand niet meer als een exchange concept gezien. Men beschouwt het nu als een element dat door zijn context wordt bepaald en nog altijd gedefinieerd wordt als een collectie van assets met een herkenbare collectieve identiteit. De Items Group entiteit wordt meer naar voor geschoven waardoor de logische structuur waarop het P_META model gebaseerd is, er als volgt uitziet:



Aangezien XML een prominente rol speelde bij de creatie van P/Meta 2.0, introduceert de nieuwe standaard een aantal conventies die de overstap van P/Meta 1.2 naar XML mogelijk maken. Zo zijn alle P/Meta Attributes XML elementen of attributen geworden en is de

naamgevingconventie aangepast: i.p.v. de eerder aangehaalde Set 'S12 PERSON_DETAILS' wordt in de huidige standaard het element 'PersonDetails' van het type 'pmeta:PersonDetailsType' gebruikt. Het element 'PersonDetails' bevat o.a. de elementen PersonLastName en PersonFirstName van het type 'string' (tot en met versie 1.2 gedefinieert als de Attributes 'A89 PERSON_LAST_NAME' en 'A88 PERSON_FIRST_NAME') en het element 'Address' van het type 'pmeta:AddressType' (vroeger de set 'S12 PERSON_DETAILS').

4.2.4 SMEF-DM

De Standard Media Exchange Framework (SMEF) Data Model biedt een set van datadefinities voor alle informatie met betrekking tot productie, ontwikkeling, gebruik en beheer van media assets. De bedoeling is te verzekeren dat systemen kunnen samenwerken en informatie kunnen uitwisselen door middel van een gemeenschappelijk raamwerk voor de gedeelde data.

SMEF is ontwikkeld door de Media Data Goup (BBC Technology) in opdracht van de BBC. Dit model is afgewogen tegen de relevante standaarden zoals MPEG-7, P/META, ISAN/V-ISAN en TV Anytime. Allen hebben ze bijgedragen tot de definiëring van het SMEF datamodel.

Het SMEF datamodel is gebaseerd op metadata die geassocieerd zijn met media. Het bereik van de metadata gaat verder dan de beschrijvende metadata. In het model zitten de nodige data om businessprocessen te ondersteunen, van commissie en het vastleggen van video tot transmissie en archivering. Het model is bruikbaar in de sector van televisie, radio en het web en biedt ondersteuning aan zowel analoge als digitale services.

Het SMEF datamodel wordt gebruikt als centrale bron van datadefinities voor de ontwikkeling van applicaties voor de BBC. Het model biedt een initiële set van datadefinities voor de projecten en wordt doorheen het project aangewend om de naleving te beoordelen van de gebruikte datadefinities in het project aan de SMEF standaard.

Het model is ontwikkeld voor gebruik binnen de BBC. Toch zijn de datadefinities voldoende generisch om bruikbaar te zijn buiten de BBC-context. Sommige referenties in het SMEF datamodel zijn echter heel specifiek voor de BBC, zoals het World Service Programme Numbers. Zo heeft de BBC ook een specifieke interpretatie van het begrip "programme week" (vervat in de entiteit PRGRAMME_WEEK_CALENDAR_YEAR). De gebruikers buiten

de BBC-omgeving moeten hier echter van op de hoogte zijn, hoewel het model ontwikkeld is om de media asset management en broadcasting-wereld als geheel te beschrijven.

Zoals reeds eerder vermeld bestaan er relaties tussen SMEF, Dublin Core en P_META. Deze relaties worden hier bondig besproken. Dublin Core is een metadatastandaard die naast andere metadatastandaarden gebruikt wordt, zoals SMEF. Hoewel dit niet specifiek is opgenomen in de documentatie rond SMEF DM, is het relatief eenvoudig om van de SMEF attributen een equivalente Dublin Core descriptor te vinden.

P_META en SMEF zijn hoogst compatibel met elkaar. P_META richt zich vooral op de business-to-business uitwisseling van programma-gerelateerde informatie en data. SMEF wordt daarentegen meer gebruikt voor de interne informatiesystemen. Het is mogelijk om directe linken te leggen tussen P_META en SMEF attributen. Notities hierover zijn opgenomen in de ducumentatie van de SMEF DM standaard.

Het SMEF datamodel bevat te veel entiteiten en relaties om leesbaar te blijven wanneer men het model in één diagram probeert voor te stellen. Het model is daarom in de documentatie onderverdeeld in acht diagrammen. Deze acht diagrammen zijn complementair en moeten beschouwd worden als een “window” op het hele datamodel. Een individuele entiteit kan dus in verschillende diagrammen voorkomen. De acht diagrammen behandelen de volgende concepten: editoriaal object, media object, materiaal instantie, subject en referentie, commissie, editoriaal genre en beschrijving, contract en rol, publicatie en ten slotte publiek.

SMEF is een heel groot model en moeilijk in één keer te begrijpen. Om te vermijden dat de gebruikte terminologie verschillende betekenissen heeft in verschillende contexten, lijkt de gebruikte terminologie niet altijd even intuïtief. Het onderstaande overzicht zal de gebruikte terminologie verduidelijken.

Een editoriaal object in SMEF is de naam voor een volledig programma of item. Andere namen voor een editoriaal object zijn bijvoorbeeld een “werk” of een “episode”. De term editoriaal object kan in een aantal verschillende entiteiten voorkomen:

- EDITORIAL_OBJECT_GROUP: deze representeert elke groep van programma`s/items voor promotie- of verkoopsdoeleinden.
- EDITORIAL_OBJECT_CONCEPT: geeft een beschrijving van de eigenschappen van één enkel programma/werk dat algemeen op alle versies van toepassing is.
- EDITORIAL_OBJECT_VERSION: is de beschrijving van een versie van een editoriaal object voor een specifiek doeleinde.

Het Image Format Type definieert de geometrische eigenschappen van een beeld of beeldapparaat. De BBC introduceerde een Publicatie Formaat Code om deze informatie voor te stellen. Deze code bestaat uit zes karakters en heeft het volgende formaat: aabccd waarbij:

- aa= Active Image Aspect Ratio
- b = Display Format
- cc= Raster Aspect Ratio
- d = Protected Aspect Ratio

De codes voor aa en cc kunnen de volgende waarden hebben:

- 16 = 16:9
- 15 = 15:9
- 14 = 14:9
- 12 = 12:9 = 4:3

De code b kan de volgende waarde hebben:

- P = Pillarbox
- L = Letterbox
- F = Full Frame
- M = Mixed Formats

Het Display Format kan in feite worden afgeleid uit de drie andere parameters, maar wordt toch expliciet opgenomen in de code.

Het ACQUISITION_BLOCK is een groep van audio items in een gepubliceerde vorm, zoals een CD of record. Een set acquisition blocks vormt een catalog van vooraf opgenomen muziek en spraakitems. Op zich betreft het hier niet steevast CD`s, maar veeleer een aanduiding dat zo`n set van records beschikbaar is. Een voorbeeld hiervan is “The Bes of Des O`Connor”. MUSIC_SPEECH_SOUND_ITEM_IN_BLOCK is een individueel item van een acquisition block. Dit kan bijvoorbeeld een track zijn van een CD. Een voorbeeld hiervan is de track “Moon River” van de collectie “Best of Des O`Connor”. De beschrijvende informatie omtrent dit item wordt beschreven door één van de subelementen van EDITORIAL_OBJECT_VERSION.

Een Media Object is de beschrijving van een component van een editoriaal object. Het kan bijvoorbeeld de audio, video en ondertiteling voorstellen. De entiteit MEDIA_OBJECT representeert de metadata betreffende algemene en editoriale informatie van een media

object zoals de editoriale beschrijving van een audio clip. UNIQUE_MATERIAL_INSTANCE bevat de attributen die de opslag van een Media Object beschrijven. Eén van de attributen van een unieke materaalinstantie is UMID (Unique Material Identifier) die SMPTE standaard volgt voor de identificatie van materiaal. MEDIA_OBJECT_GROUP legt de editoriale en conceptuele verbanden vast tussen mediaobjecten. Een voorbeeld hiervan is een set van specifieke audioclips die samen gegroepeerd zijn als een tussenstap in het productieproces. De entiteiten STORAGE en STORAGE_TYPE duiden aan waar en in welke vorm het unieke materiaal wordt bewaard.

De groep Location, Story en Classification geven een samenvatting van de verschillende manieren waarop individuele programma`a en items worden gecatalogeerd en geclassificeerd. De locatie (CONTENT_LOCATION) set geeft de locaties aan die betrokken zijn in een media asset. STORY duidt het thema aan die van toepassing is op de verschillende versies van programma`s of items.

Een andere groep entiteiten beschrijft gegevens met betrekking tot transmissie en publicatie van editoriale objecten. De kernentiteit hiervan is PUBLICATION_EVENT die de geplande en actuele transmissie en/of publicatie van een programma representeert. Een publicatie event kan worden onderverdeeld in andere publicatie events. Dat is handig voor publicatie events die in feite een container vormen en verder kunnen worden onderverdeeld in andere publicatie events.

Mensen en organisaties worden beschreven door een andere groep entiteiten. Zij kunnen verschillende rollen uitoefenen in de media asset management zoals een cameraman, een director of een persoon die de rechten bezit op een gebruikte locatie. SMEF ondersteunt de verschillende rollen die een persoon of organisatie in deze context kunnen vervullen. De entiteit PERSON reflecteert een belangrijke actor in de sector. De entiteit ORGANISATION behandelt groepen zoals bijvoorbeeld een erkend, onafhankelijk productiehuis. De entiteit PERSON_LINK_ORGANISATION ondersteunt de relaties die personen en organisaties kunnen hebben. De entiteit ROLE beschrijft dan de taak of verantwoordelijkheid van de persoon en/of organisatie. De associatie kan een contract inhouden en wordt gerepresenteerd door de entiteit CONTRACT en CONTRACT_LINE.

4.2.5 MARC/MARC21

MARC is een acroniem van Machine-Readable Cataloging. MARC is een standaard voor de representatie en de communicatie van bibliografische en aanverwante informatie en dit in een machine-leesbare vorm, aangevuld met aanverwante documentatie. De standaard wordt onderhouden door de Amerikaanse Library of Congres en vindt zijn oorsprong in de jaren 1960 als een digitale vorm van bibliotheekfiches. De hoofdfunctie van de standaard was dan ook de vereenvoudiging en bespoediging van het terugvinden van boeken in de bibliotheek. De MARC data-elementen vormen dan ook de basis voor de meeste bibliotheekcatalogi. Verder is er geen enkel alternatief met een gelijkaardige graad van gegranuleerdheid.

MARC ondersteunt acht soorten materiaal waaronder het type “Sound recordings” dat alle soorten geluid omvat met uitgezondering van muziek. Hieronder vallen dus ook mondelinge historische bronnen. Verder bevat MARC ook zeven types records waaronder “Computer file”, zoals een gedigitaliseerde versie van een mondelinge historische bron, en “Manuscript (textual) language material”, zoals de transcriptie van een mondelinge historische bron.

Een MARC (bibliografisch) record bestaat uit meerdere velden. Er zijn velden voor auteur, titelinformatie, enzovoort. Deze velden kunnen verder worden onderverdeeld in subvelden.

De tekstuele namen van de velden (zoals auteur en onderwerp) worden vervangen door tags die bestaan uit een driecijferige code. Deze code beschrijft dus welke gegevens in het veld staan.

Subvelden worden gescheiden door middel van een karakter (bvb $) dat aangevuld wordt met een subveldcode die aangeeft welke gegevens volgen.

Sommige velden worden verder gedefinieerd door indicatoren. Dit zijn 2 posities die een karakter tussen 0 en 9 kunnen bevatten. Het 2de karakter kan bijvoorbeeld aangeven dat een aantal volgende karakters door de computer moet worden genegeerd bij de sortering. Dit kan bijvoorbeeld gebruikt worden bij auteurs met een familienaam die met “van” begint.

Een eenvoudig voorbeeld van een MARC Entry:

245 10 $a Interview met een oudstrijder**$h**[sound recording].

260 ## $a Kortrijk**$b**Vereniging voor oudstrijders**$c**1999.

300 ## $a**1 minidisc**$bdigital, ATRAC, stereo.

500 ## $a Interview met een oudstrijder 1940-1944.

500 ## $atranscriptie beschikbaar.

511 0# $a Interview afgenomen door X

De eerste regel bevat een veld met de code 245 die wijst op een “Title Statement”. De indicatoren hebben de waarde 1 en 0 en het veld bevat de subvelden $a, de eigenlijke titel, en $h het medium.

Om de records overzichtelijker te maken en bewerking van de records te vereenvoudigen is later de MARC XML-standaard ontworpen die MARC-records in XML-bestand voorstelt.

Het voorbeeld toont de hoge mate van gegranuleerdheid die het MARC-formaat biedt en de daarmee gepaard gaande complexiteit. Het formaat is desondanks compact. Veldnamen zoals “plaats van publicatie” worden immers vervangen door een korte code.

Bovendien zit er een logica in de veldcodes, wat de complexiteit iets vermindert (bvb 6XX betekent veld met informatie over het onderwerp, X00 betekent een naam).

MARC kent geen semantische zoekfunctie. Dit wil zeggen dat gezocht wordt naar de gegeven sleutelwoorden in de verschillende velden maar dat geen rekening wordt gehouden met de betekenis of het concept van de sleutelwoorden.

Naast de bibliografische records die de kenmerken van resources bespreken zijn er nog andere types records die bijvoorbeeld een classificatie beschrijven of informatie geven over namen, onderwerpen, enzovoort.

Nadelen MARC/MARC21:

- Complex
- Geen hiërarchische opbouw
- Geen semantiek
- Niet geschikt voor “leken”

Voordelen MARC/MARC21:

- Een hele hoge gegranuleerdheid
- Wijdverspreid
- Kunnen in XML worden weergegeven

4.2.6 MODS

Het Metadata Object Description Scheme (MODS) is een schema voor een bibliografische set van elementen die kan gebruikt worden in een breed gamma van toepassingen, voornamelijk bibliotheektoepassingen. De standaard wordt onderhouden door de Network Development and MARC Standards Office met input van de gebruikers. MODS wordt beschreven in XML. Als XML-schema moest de MODS-standaard data kunnen bevatten van bestaande MARC-21-bestanden. MODS bevat een subset van MARC-velden en maakt gebruik van taalgebaseerde tags in plaats van numerieke tags zoals in MARC. In sommige gevallen hergroepeert de standaard elementen van de MARC-21-standaard.

MODS kan gebruikt worden in de volgende toepassingsgebieden:

- Als een SRU-formaat
- Als een uitbreidingsschema voor METS
- Om metadata voor te stellen bij het harvesten
- Als een manier om bronnen te beschrijven in XML
- Om een vereenvoudigd MARC-record voor te stellen in XML

De bedoeling van MODS is om andere metadataformaten te complementeren. Voor sommige toepassingen, zeker toepassingen die MARC gebruiken, zijn er verschillende voordelen van MODS ten opzichte van andere schema`s. Zo is de set elementen van MODS rijker dan die van Dublin Core en eenvoudiger dan het volledige MARC-formaat. MODS-elementen zijn bovendien compatibeler met data uit bibliotheken dan ONIX. Het schema dat MODS gebruikt, is meer gericht op de eindgebruiker dan bijvoorbeeld MARC, die numerische tags gebruikt.

Naast deze voordelen kent MODS nog enkele features die interessant kunnen zijn:

- De elementen erven de semantiek van MARC.
- Sommige data worden anders gegroepeerd: wat bij MARC in verschillende data-elementen wordt beschreven kan in MODS soms in één data-element worden vervat.
- MODS veronderstelt niet het gebruik van één of andere cataloguscode zoals bij MARC het geval is.
- Verschillende elementen hebben optioneel een ID-attribuut waardoor gemakkelijk links gelegd kunnen worden op elementniveau.

Zoals reeds eerder aangegeven heeft MODS een subset elementen overgenomen van MARC-21. Als een elementset die de representatie toelaat van data die reeds in MARC zijn beschreven, heeft deze standaard de conversie van de core-elementen tot doel terwijl de meer specifieke data genegeerd kunnen worden. Het MODS-schema beoogt geen round-tripability met MARC-21. Dat wil zeggen dat een MARC-21-record naar MODS geconverteerd kan worden maar niet terug naar het originele MARC-21-record.

MODS wordt geserialiseerd in XML. Hiertoe definieert MODS hoofdelementen, subelementen en attributen van de hoofd- en subelementen. De inhoud van de elementen wordt ingevuld op het laagste niveau om op deze manier “mixed elements” te vermijden. Als bijvoorbeeld <titleInfo> enkele subelementen bevat voor <title>, <partNumber> en <partName>, dan is <titleInfo> slechts een wrapper-tag die de meer specifieke elementen <title>, <partNumber> en <partName> bevat.

Attributen kunnen op elk niveau geassocieerd worden met elementen en worden gedefinieerd met het element waarmee ze geassocieerd worden. Enkele gebruikelijke attributen zijn: type, encoding en authority.

Een MODS-document heeft een schemadeclaratie die de namespace aangeeft. Binnen een record of groep van records is deze schemadeclaratie voor elk element optioneel aangezien de MODS namespace binnen het record wordt aangegeven. Het is echter heel gebruikelijk om het prefix “mods:” voor elk element te gebruiken wanneer een MODS-record wordt gecombineerd met XML-data van een andere namespace: bijvoorbeeld een MODSrecord binnen een METS-document.

In MODS zijn er geen verplichte elementen. De enige voorwaarde die hiervoor wordt opgelegd is dat een MODS-beschrijving minstens één element bevat.

Hieronder wordt een voorbeeld gegeven van een MODS-beschrijving van een hoofdstuk uit een boek:

```
<mods xmlns:xlink="http://www.w3.org/1999/xlink" version="3.0"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xmlns="http://www.loc.gov/mods/v3"
xsi:schemaLocation="http://www.loc.gov/mods/v3
http://www.loc.gov/standards/mods/v3/mods-3-0.xsd">
  <titleInfo>
  <title>Models, Fantasies and Phantoms of Transition</title>
  </titleInfo>
  <name type="personal">
  <namePart type="given">Ash</namePart>
```

```
  <namePart type="family">Amin</namePart>
  <role>
  <roleTerm type="text">author</roleTerm>
  </role>
  </name>
  <typeOfResource>text</typeOfResource>
  <relatedItem type="host">
  <titleInfo>
  <title>Post-Fordism</title>
  <subTitle>A Reader</subTitle>
  </titleInfo>
  <name type="personal">
  <namePart type="given">Ash</namePart>
  <namePart type="family">Amin</namePart>
  <role>
  <roleTerm type="text">editor</roleTerm>
  </role>
  </name>
  <originInfo>
  <dateIssued>1994</dateIssued>
  <publisher>Blackwell Publishers</publisher>
  <place>
  <placeTerm type="text">Oxford</placeTerm>
  </place>
  </originInfo>
  <part>
  <extent unit="page">
  <start>23</start>
  <end>45</end>
  </extent>
  </part>
  </relatedItem>
  <identifier>Amin1994a</identifier>
</mods>
```

4.2.7 CDWA

Categories for the Description of Works of Art (CDWA) beschrijft de data uit kunstdatabanken aan de hand van een conceptueel raamwerk voor het beschrijven en opvragen van informatie over kunstwerken, achitectuur of ander cultureel materiaal. Het CDWA bevat 512 categorieën en subcategorieën. Een kleine subset van deze categorieën vormt de core. Deze categorieën stellen de minimale informatie voor die nodig is om een werk te beschrijven en te identificeren.

Daarbuiten behelst de CDWA ook discussies, basisregels voor het catalogiseren en voorbeelden.

Het CDWA is een product van de ART Information Task Force (AITF), die de dialoog aanzwengelt tussen kunsthistorici, kunstinformatie-professionals en informatieleveranciers zodat ze samen richtlijnen kunnen ontwikkelen voor de beschrijving van kunstwerken, architectuur, groepen van objecten en visuele en tekstuele surrogaten.

Deze groep is opgericht begin jaren `90 en bestond uit vertegenwoordigers van verschillende gemeenschappen die kunstinformatie aanleveren en gebruiken: museumcurators, professionals op het gebied van visuele bronnen, kunstbibliothecarissen, informatiemanagers en technische specialisten.

De categorieën leveren een raamwerk waarop bestaande informatiesystemen kunnen gemapt worden en op basis waarvan nieuwe systemen ontwikkeld kunnen worden. Daarbij identificeren de discussies in het CDWA woordenschatten en beschrijvende toepassingen die de informatie in de verschillende systemen meer compatibel en meer toegankelijk maken.

Het gebruik van het CDWA raamwerk moet bijdragen tot de integriteit en de levensduur van de data en zal de migratie van de data naar nieuwe systemen in de toekomst vergemakkelijken. Bovenal zal het de eindgebruiker helpen in het opzoeken van betrouwbare informatie, ongeacht het systeem waarin de data zijn opgeslagen.

Het CDWA stelt een relationele datastructuur voor, waar records over objecten of werken aan elkaar gelinkt zijn met hiërarchische relaties. Het CDWA raadt ook aan om aparte files bij te houden voor gerelateerde visuele werken, gerelateerd tekstueel materiaal, personen en bedrijveninformatie, lokaties en dergelijke. Authoriteitsinformatie over personen, plaatsen, concepten en andere onderwerpen kunnen belangrijk zijn om informatie terug te vinden, maar deze informatie wordt beter in afzonderlijke files bijgehouden dan geïntegreerd in de informatie over een werk. Het voordeel hiervan is dat deze informatie maar eenmaal beschreven moet worden en gebruikt kan worden in de records over kunstwerken.

Het CDWA Lite is een XML-schema dat de core-elementen bevat voor de beschrijving van een kunstwerk of cultureel materiaal en dat gebaseerd is op het CDWA en het CCO. CDWA Lite-records willen bijdragen aan het verenigen van catalogi en bibliotheken die gebruik maken van het OAI-PMH-protocol voor de uitwisseling van data.

Het CCO (Cataloging Cultural Objects: A Guide to Describing Cultural Works and Their Images) levert een aantal voorgeschreven richtlijnen aan voor de selectie, ordening en formattering van data die gebruikt worden om catalogusrecords aan te vullen. Het gebruikt

hiervoor informatie die gerelateerd is aan een subset van de CDWA-categorieën en de VRA Core-categorieën.

4.2.8 VRA Core

De VRA Core (Visual Resources Association Data Standards Committee) is een datastandaard voor het beschrijven van culturele erfgoed. Het bestaat uit een set van metadata elementen en hoe deze elementen moeten worden gestrucureerd. Deze set elementen categoriseert de beschrijving van de werken alsook de beelden die het werk afbeelden.

De Visual Resources Association Data Standards Committee heeft de core categorieën aangepast aan de laatste ontwikkelingen in data standaarden. Op dit moment zit men aan de vierde editie Core 4.0

De veranderingen waren nodig om een xml-representatie van de core te kunnen maken. De veranderingen betreffen vooral de herdefinitie van wat in versie 3.0 gekend was als element qualifiers. De qualifiers zijn geconverteerd naar subelementen en attributen volgens de xml-encoding syntaxis.

Een andere verandering is het type dat aan een record kan toegewezen worden. Vroeger was dit work en image. Nu is daar bij de vierde release van het schema nog een derde type bijgekomen: collection. Een object van het type work is een unieke entiteit zoals een object of een event. Voorbeelden hiervan zijn schilderijen, beeldhouwwerken of een voorstelling. Een record van het type image is een visuele representaie van een work als zijn geheel of een deel ervan. Een collectie is een aggregatie van work of image records. Dit was nodig voor het catalogiseren op collectie-niveau toe te laten.

Het enige element in een VRA Core 4.0 record dat noodzakelijk is, is die informatie die een record enkelvoudig kan identificeren. In non-XML formaat betekent dit dat er minstens een work, collection of image element in het record moet aanwezig zijn. In een XML formaat is dit de wrapper die deze informatie bevat.

Het is echter aangewezen om nog extra beschrijvende elementen op te nemen in een record om deze snel en gemakkelijk terug te vinden. Dit zijn de elementen die informatie leveren over de fundamentele vragen over het object: wat, wie, waar en wanneer. Hiertoe zijn de volgende elementen aangewzen om ook steeds op te nemen in de beschrijving van een object:

```
WORK TYPE (wat)
```

```
TITLE (wat)
AGENT (wie)
LOCATION (waar)
DATE (wanneer)
```

VRA Core 4.0 baseerd zich op richtlijnen gegeven door het CCO. Deze richtlijnen postuleren dat er rekening moet gehouden worden met display en indexing requirements. Dit wil zeggen dat data waarden voor een gegeven element zouden geformateerd moeten zijn in de vorm dat ze getoond worden. Tegelijkertijd moeten al deze data waarden apart geformateerd worden en moeten ze gelinkt zijn aan thesauri om het terugvinden van de data te vergemakkelijken.

Hiertoe heeft het VRA Core XML schema elk element voorzien van twee subelementen, display en notes. Deze subelementen worden genest binnen de wrapper van het element:

```
<materialSet>
        <display>oil on canvas</display>
        <notes source=”Art Bulletin, v.87, no.1 (March 2005)”>Medium
        originally thought to be tempera. Oil medium discovered in tests at
        Uffizi in 2003</notes>
        <material type=”medium” vocab=”AAT” refid=”300015050”>oil
        paint</material>
        <material type=”suppert” vocab=”AAT”
        refid=”300014078”>canvas</material>
</materialSet>
```

Het optionele notes subelement binnenin de wrapper laat toe om vrije tekst annotaties toe te voegen die nog niet is gecovered door de andere attributen. Als de annotatie echter verwijst naar het gehele record, dan moet het description element gebruikt worden.

Wat in de vorige versie van de VRA Core 3.0 gekend was als het RECORD TYPE werd in deze laatste versie herwerkt tot een element WORK, COLLECTION of IMAGE. Dit element wordt gebruikt als opslagplaats voor de administratieve data. Hiervoor werden twee attributen voor deze elementen aangeleverd:

- Id: deze bevat een unieke identifier van het XML record
- Het globale refid attribuut: dit attribuut kan gebruikt worden om een lokaal nummer, code of adres op te slaan dat het record op een unieke manier identificeert binnen de context van het source attribuut.

- Het globale source attribuut: deze slaat de set of de omgeving op waartoe het record behoort, zoals een museum.

In het XML schema functioneren de functies Work, Image of Collection als een algemene wrapper waarbinnen de andere elementen, elk verpakt in sets, zich bevinden. De attributen id, refid en source van een Work, Image of Collection wrapper dienen als een unieke identificatie in verschillende contexten. Het id attribuut identificeert een XML record binnen een bestand dat vele XML records bevat. Refid en source identificeren een XML record binnen het systeem, vanwaar ze afkomstig is. Hieronder staat een voorbeeld:

```
<work id="w_98765432" refid="14363" source="History of Art Visual Resources
Collection, UCB">
        <agentSet>
                <display></display>
                <notes></notes>
                <agent></agent>
        </agentSet>
        <dateSet></dateSet>
        <culturalContextSet></culturalContextSet>
        <descriptionSet></descriptionSet>
        …
</work>
```

Op dit moment zijn er twee XML schema`s van de VRA Core 4.0 metadatastandaard: restricted en unrestricted. De unrestricted versie legt geen beperkingen op aan de datawaarden die worden ingevuld in een VRA Core 4.0 record, terwijl de restricted versie dit wel doet.

Er dient nog gezegd worden dat het doel van de VRA Core 4.0 file sharing is. Dit wil zeggen dat de standaard tekort schiet om alle data in XML op te slaan. Deze standaard moet waarschijnlijk nog uitgebreid worden met andere sub-elementen en attributen om aan de noden van een organisatie te voldoen. Als uitwisselingsschema is de VRA Core 4.0 wel voldoende.

4.2.9 EAD

EAD is een acroniem voor Encoded Archival Description en is een metadatastandaard die is ontwikkeld door de bibliotheek van de University of Berkeley in Californië. Deze wou meer informatie kunnen invoeren dan die voorzien bij MARC records.

Tot de vereisten behoorden:

- Mogelijkheid tot het weergeven van uitgebreide en intergerelateerde beschrijvende informatie
- Mogelijkheid tot het bewaren van de hiërarchische relaties tussen verschillende niveaus van beschrijving
- Mogelijkheid tot het weergeven van beschrijvende informatie die geërfd wordt over de hiërarchische niveaus heen
- Mogelijkheid te navigeren binnen een hiërarchische informatiestructuur
- Ondersteuning voor elementspecifieke indexering en navigatie.

EAD is SGML-gebaseerd maar ondersteunt ook XML. De elementen die mogen gebruikt worden om een manuscript collectie te beschrijven en de ordening van deze elementen (bvb. welke elementen nodig zijn, welke elementen toegelaten zijn binnen andere elementen) worden gespecificeerd in de EAD Document Type Definition (DTD). De tag set hierin gespecificeerd bestaat uit 146 elementen en wordt gebruikt zowel voor het beschrijven van een collectie in zijn geheel als voor de encodering van een gedetailleerde multi-level inventaris van de collectie. Vele EAD elementen zijn, of kunnen worden, afgebeeld op andere standaarden zoals MARC en Dublin Core, wat de flexibiliteit en interoperabiliteit van de data verhoogd.

Een EAD bestaat uit verschillende onderdelen:

- De EAD-header bevat de titel en gedetailleerde informatie over de collectie en het document. De elementen in de header worden vaak ook gemapt naar Dublin Core-elementen
- De archiefbeschrijving bestaat uit de Data Item Description (DID), aangevuld met eventuele extra beschrijvingen en vervolgens, het grootste deel, de volledige inventaris van de collectie

De did bevat een beschrijving van de collectie in zijn geheel, inclusief de beheerder (persoon of organisatie), taal, korte beschrijving,… Deze DID kan gevolgd worden door verschillende extra elementen:

- Een biografische beschrijving van de persoon of organisatie
- Een uitgebreide beschrijving van de collectie
- Beschrijving van objecten gerelateerd aan de collectie

- Objecten die tot de collectie behoren maar die gescheiden zijn van de collectie (bvb voor speciale behandeling, specifieke opslagbehoeften,…)
- Een lijst van onderwerpen of trefwoorden voor de collectie
- Beperkingen op het materiaal in de collectie.

De inventaris van de collectie wordt progressief opgedeeld in kleinere stukken met steeds "fijnere" informatie. Dit laat toe om bij zoekopdrachten en inventariseren de gewenste informatiediepte te bepalen.

Verder biedt The Research Libraries Group een "coördinatiecentrum" aan. Leden kunnen hun informatie doorgeven aan deze groep en deze zal dan de gegevens indexeren en een zoekinterface genereren voor deze index. Dit laat onderzoekers toe met één enkele query te zoeken in honderden collecties.

Een voorbeeld van een EAD-bestand:

```
<filedesc>
    <titlestmt>
        <titleproper> Interview met een oudstrijder
               <date>1999</date>
        </titleproper>
        <author>Vereniging voor oudstrijders</author>
    </titlestmt>
    <notestmt>
        <note>
        <p>
            <subject>Wereldoorlog II</subject>
        </p>
    </note>
    </notestmt>
</filedesc>
```

Voordelen EAD:

- Ondersteunt hiërarchie
- Kan worden vertaald naar MARC en Dublin Core

Nadelen EAD:

- SGML is minder gebruiksvriendelijk

4.2.10 SPECTRUM

SPECTRUM is een open standaard die wordt onderhouden en gestuurd door MDA. Deze standaard die procedures beschrijft voor het documenteren, de behandeling en de

identificatie van objecten. Verder wordt er ook aandacht besteed aan bvb. rechtenbeheer, uitleenbeheer en risicobeheer. SPECTRUM evolueert continu en biedt dan ook de mogelijkheid tot groei en uitbreiding.

SPECTRUM is een Britse standaard die werd ontwikkeld met de hulp van de ervaring en het inzicht van honderden personen uit de museumbranche. Deze standaard wordt dan ook gezien als de “industriestandaard” voor documenteren.

4.2.11 ISAD(G)

De standaard binnen de archiefsector is ISAD(G) of General International Standard Archival Description. Deze standaard moet helpen bij het opstellen van beschrijvingen van collecties en objecten. De standaard bestaat uit verschillende regels die moeten gevolgd worden. Voor multi-level beschrijvingen (zoals mogelijk bij EAD) wordt bvb. aangeraden de beschrijvingen van algemeen naar bijzonder in te vullen en de plaats van een beschrijvingseenheid in de hiërarchie duidelijk te maken.

Op gelijkaardige manier worden ook regels gegeven voor het invullen van referenties, titels, datering, enz.

Een standaard afgeleid van ISAD(G) is SEPIADES (SEPIA Data Element Set). Deze standaard is gericht op het beschrijven en beheren van fotografische collecties en bevat 21 “Core”-elementen, aangevuld met meer dan 400 andere elementen. SEPIADES gebruikt een multi-level aanpak gelijkaardig aan ISAD(G). De standaard voorziet niet in een eigen encodering en is dus eerder een “handleiding” voor het beschrijven van collecties. Voor het opslaan van records wordt Dublin Core aangeraden.

4.2.12 ISAAR

Deze norm verschaft richtlijnen voor het maken van archivistische geautoriseerde beschrijvingen van entiteiten (organisaties, personen en families) betrokken bij de vorming en het beheer van archieven.

Archivistische geautoriseerde beschrijvingen kunnen worden gebruikt:

- om een organisatie, persoon of familie als eenheid binnen een archivistisch beschrijvingssysteem te beschrijven; en/of

- om de creatie en het gebruik van ontsluitingstermen in archivistische beschrijvingen te regelen;
- om de relaties te documenteren tussen verschillende archiefvormers en tussen deze entiteiten en de archieven die zij gevormd hebben en/of andere bronnen over of van hen.

Het beschrijven van archiefvormers is een wezenlijke taak van archivarissen, ongeacht of de beschrijvingen zich in papieren of digitale systemen bevinden. Dit vereist een volledige en geactualiseerde documentatie van de context van archiefvorming en -gebruik, vooral van de herkomst van de archieven.

De parallelnorm van dit document, ISAD(G): Algemene Internationale Norm voor Archivistisch Beschrijven, voorziet in het opnemen van contextgegevens in archivistische beschrijvingen op elk mogelijk niveau. ISAD(G) onderkent ook de mogelijkheid om contextgegevens afzonderlijk vast te leggen en te onderhouden, en om deze contextgegevens te koppelen aan de andere gegevenselementen gebruikt om archieven te beschrijven.

Er zijn meerdere argumenten waarom het afzonderlijk vastleggen en onderhouden van dit type contextgegevens een essentieel onderdeel van het archivistisch beschrijven is.

Deze werkwijze maakt het mogelijk om beschrijvingen van archiefvormers en contextgegevens te koppelen aan beschrijvingen van archiefstukken van diezelfde archiefvormer(s) die mogelijk in verschillende archiefbewaarplaatsen berusten, alsmede ze te koppelen aan beschrijvingen van andere bronnen zoals bibliotheekmateriaal en museumobjecten die met de entiteit verband houden. Zulke relaties bevorderen de uitvoering van het archiefbeheer en ondersteunen het onderzoek.

Archiefbewaarplaatsen die archiefstukken van eenzelfde bron beheren, kunnen de contextgegevens over deze bron eenvoudiger delen of ernaar verwijzen als die op een gestandaardiseerde manier zijn beheerd. Zo'n standaardisatie is vooral van belang in een internationale context aangezien het delen of koppelen van contextgegevens over nationale grenzen heen kan gaan. Het multinationale karakter van archivering, zowel nu als in het verleden, vormt een stimulans tot internationale standaardisatie die dan weer de uitwisseling van contextgegevens zal bevorderen. Processen zoals kolonisatie, immigratie en handel hebben bijvoorbeeld bijgedragen aan het multinationale karakter van archivering.

Deze norm stimuleert de creatie van consistente, geschikte en duidelijke beschrijvingen van archiefvormende organisaties, personen en families, met als doel het gemeenschappelijk gebruik van archivistische geautoriseerde beschrijvingen te bevorderen. Hij moet worden gebruikt in combinatie met bestaande nationale normen of als basis voor de ontwikkeling van nationale normen.

Archivistische geautoriseerde beschrijvingen zijn vergelijkbaar met bibliografische geautoriseerde beschrijvingen aangezien beide de creatie van gestandaardiseerde ontsluitingstermen horen te ondersteunen. De naam van de archiefvormer van de beschrijvingseenheid is één van de belangrijkste ontsluitingstermen. Ontsluitingstermen kunnen kwalificaties meekrijgen die essentieel worden geacht om de identiteit van de benoemde entiteit te verhelderen, zodat een nauwkeurig onderscheid kan worden gemaakt tussen verschillende entiteiten met een gelijke of gelijkende naam.

Archivistische geautoriseerde beschrijvingen moeten echter aan meer eisen voldoen dan bibliografische. Deze aanvullende eisen komen voort uit het belang dat men in archivistische beschrijvingssystemen hecht aan het documenteren van archiefvormers en de context van archiefvorming. Archivistische geautoriseerde beschrijvingen gaan dus veel verder en zullen gewoonlijk meer gegevens bevatten dan bibliografische.

Het belangrijkste doel van deze norm is dus om algemene richtlijnen te verschaffen voor de standaardisatie van archivistische beschrijvingen van archiefvormers en de context van archiefvorming, die het volgende mogelijk maken:

- het raadplegen van archieven, gebruik makend van beschrijvingen van de context van archiefvorming, die zijn gekoppeld aan de beschrijvingen van vaak verschillende en fysiek verspreid liggende archiefstukken;
- de context waarin archieven zijn ontstaan en gebruikt te begrijpen zodat de gebruiker inzicht verwerft in de betekenis en het belang van die archieven;
- een nauwkeurige identificatie van archiefvormers, met inbegrip van beschrijvingen van de relaties tussen verschillende entiteiten, met name het documenteren van administratieve veranderingen binnen organisaties of veranderingen in de persoonlijke omstandigheden van individuen en families;
- de uitwisseling van die beschrijvingen tussen instellingen, systemen en/of netwerken.

Een archivistische geautoriseerde beschrijving opgesteld volgens deze norm, kan ook worden gebruikt om de naam en identiteit van een organisatie, persoon of familie te beheren in een ontsluitingsterm die verbonden is met de archivistische beschrijvingseenheid.

OPBOUW EN GEBRUIK VAN DE NORM

Deze norm bepaalt welke soort gegevens kunnen worden opgenomen in een archivistische geautoriseerde beschrijving. Hij verschaft richtlijnen voor het opnemen van dergelijke beschrijvingen in een archivistisch beschrijvingssysteem. De inhoud van de gegevenselementen in een geautoriseerde beschrijving zal worden vastgesteld volgens de afspraken en regels van de instantie die haar opstelt.

Deze norm bestaat uit gegevenselementen, ieder element bestaat uit:

- de naam van het beschrijvingselement
- een formulering van het doel van het beschrijvingselement
- een formulering van de regel (of regels) die op het element van toepassing zijn
- waar relevant, voorbeelden die de toepassing van de regel illustreren.

De paragrafen zijn enkel genummerd om ze te kunnen citeren. Deze nummers moet men niet gebruiken om beschrijvingselementen aan te duiden of om de volgorde of structuur van de beschrijvingen voor te schrijven.
De beschrijvingselementen voor een archivistische geautoriseerde beschrijving zijn ingedeeld in vier velden:

- Identiteit
- Beschrijving
- Relaties
- Beheer

(met gegevens die de geautoriseerde beschrijving op een unieke wijze identificeren en met gegevens die aangeven hoe, wanneer en door welke instantie de geautoriseerde beschrijving werd gecreëerd en onderhouden).

Alle elementen kunnen worden gebruikt, maar de volgende vier elementen zijn noodzakelijk:

- Soort entiteit
- Geautoriseerde naam(namen)
- Bestaansperiode
- Identificatiecode van de geautoriseerde beschrijving

De aard van de beschreven entiteit en de eisen van het systeem of netwerk waarin de vervaardiger van een archivistische geautoriseerde beschrijving werkt, zullen bepalen welke optionele beschrijvingselementen in een bepaalde geautoriseerde beschrijving worden gebruikt en of deze elementen in een verhalende en/of gestructureerde vorm worden gepresenteerd.

Veel van de beschrijvingselementen in een volgens ISAAR(CPF) opgestelde geautoriseerde beschrijving zullen worden gebruikt als ontsluitingstermen. Regels en afspraken over de standaardisatie van ontsluitingstermen kunnen zowel nationaal als per taal worden ontwikkeld. De woordenlijsten en afspraken voor het opstellen of selecteren van de inhoud van deze elementen kunnen ook nationaal of per taal worden ontwikkeld.

De voorbeelden in deze norm zijn illustratief, het zijn géén voorschriften. Zij zijn bedoeld om de regels waar ze bij horen te verhelderen, niet om ze uit te breiden. Beschouw noch de voorbeelden, noch de vorm waarin ze worden gepresenteerd als instructies. Om de context van elk voorbeeld duidelijk te maken, wordt het gevolgd door een vermelding, in cursief, van de naam van de instantie die het heeft geleverd. Om de context duidelijk te maken, wordt elk voorbeeld gevolgd door een vermelding, in cursief, van de naam van de instantie die het voorbeeld heeft geleverd. Verdere verklarende aantekeningen kunnen volgen, ook in cursief, voorafgegaan door de afkorting N.B.: Verwar de bronverwijzing van het voorbeeld of mogelijke aantekeningen niet met het voorbeeld zelf.

Deze norm is bedoeld om gebruikt te worden samen met ISAD(G) – Algemene Internationale Norm voor Archivistisch Beschrijven, 2de uitgave, en met nationale archivistische beschrijvingsnormen. Als deze normen samen worden toegepast binnen een archivistisch beschrijvingssysteem of netwerk, dan zullen geautoriseerde beschrijvingen met beschrijvingen van archieven en omgekeerd worden verbonden. Beschrijvingen van archieven kunnen verbonden worden met archivistische geautoriseerde beschrijvingen via de elementen Naam van de archiefvormer(s) en Institutionele geschiedenis / Biografie in een beschrijving die opgesteld is volgens ISAD(G).

Deze norm is bedoeld om samen met nationale normen en afspraken gebruikt te worden. Archivarissen kunnen bijvoorbeeld nationale normen volgen bij de keuze welke elementen wel of niet mogen worden herhaald. In veel landen eisen archivistische beschrijvingssystemen een enkele geautoriseerde naam voor een bepaalde entiteit, terwijl in andere landen het toegestaan is om meer dan één geautoriseerde naam te ontwikkelen.

Deze norm behandelt slechts een deel van de voorwaarden die nodig zijn om de uitwisseling van archivistische geautoriseerde gegevens mogelijk te maken. Succesvolle digitale uitwisseling van archivistische geautoriseerde gegevens via computernetwerken is afhankelijk van het gebruik van een geschikt communicatieformaat door de bewaarplaatsen die de informatie uitwisselen. Encoded Archival Context (EAC) is een dergelijk communicatieformaat. EAC ondersteunt de uitwisseling via het World Wide Web van archivistische geautoriseerde beschrijvingen die opgesteld zijn volgens ISAAR(CPF). EAC is een Document Type Definition (DTD) in XML (Extensible Markup Language) en SGML (Standard Generalized Markup Language).

4.3 Preservatie Metadatastandaarden

4.3.1 PREMIS

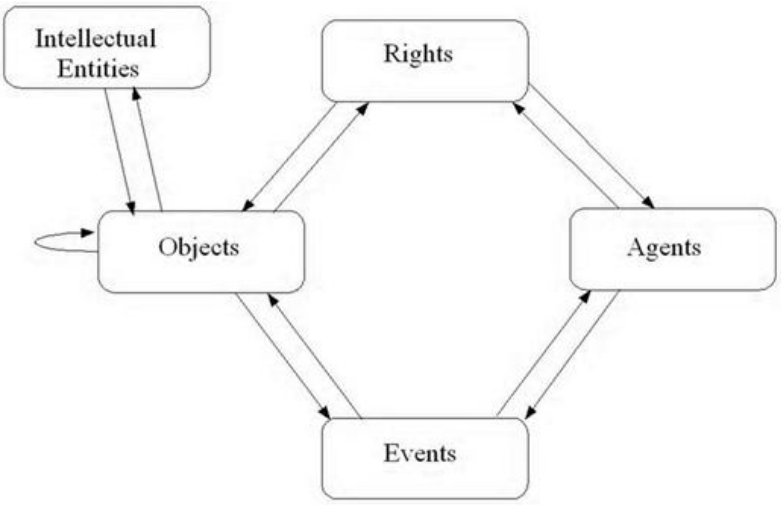
PREMIS is een preservatie metadatastandaard. Het is een project dat gestart is in juni 2003. De werkgroep bestond uit internationale experten op het gebied van preservatie en metadata. De leden kwamen uit verschillende domeinen zoals de bibliotheken, de musea, de archieven, overheidsinstellingen en de privé-sector. Het doel van deze werkgroep was de ontwikkeling van een set implementeerbare preservatie metadata die ondersteund wordt door richtlijnen en aanbevelingen voor de creatie, het beheer en het gebruik ervan. In mei 2005 werd reeds een eerste versie van hun bevindingen opgeleverd met een rapport Data Dictionary for Preservation Metadata: Final Report of the PREMIS Working Group. Dit rapport bevat tal van bronnen over preservatie metadata. Eerst en vooral is er een data woordenboek die de preservatie metadata definieert. Deze Data Dictionary is gebaseerd op het Open Archival Information System (OAIS) referentie model. Bovenop deze Data Dictionary publiceerde de werkgroep ook een set van XML schema`s die de implementatie van de Data Dictionary in digitale archiveringssystemen ondersteunt.

De PREMIS Data Dictionary definieert preservatie metadata als de informatie die een repository gebruikt om het digitale preservatieproces te ondersteunen. Meer specifiek keek de werkgroep naar metadata die de volgende functies ondersteunde:

- Uitvoerbaarheid
- Weergave
- Verstaanbaarheid
- Authenticiteit
- Identiteit

Preservatie metadata omhelst een aantal verschillende soorten metadata: administratieve metadata (met de rechten en de permissies), technische en structurele metadata. Er werd vooral aandacht besteed aan de documentatie van de oorsprong (de geschiedenis van een object) en de documentatie van de relaties, vooral deze tussen de verschillende digitale objecten.

De werkgroep ontwikkelde een simpel data model om de semantische eenheden in de Data Dictionary te definiëren. Het data model beschrijft vijf entitetiten die belangrijk zijn voor de digitale preservatie doeleinden: Intellectuele entiteiten, Objecten, Events, Rechten en Agenten zoals de onderstaande tekening verduidelijkt.



4-1: Datamodel PREMIS

- Intellectuele entiteiten: Dit is een deel van de content dat wordt beschouwd als een intellectuele eenheid met als doel het beheer en de beschrijving van de content. Bv.een boek, foto, map of databank.
- Object: een discrete eenheid van informatie in digitale vorm
- Event: een actie die een impact heeft op een object of agent
- Agent: persoon, organisatie of software applicatie die gerelateerd is aan een event van een object of geassocieerd is met de rechten van dat object.
- Rechten: beschrijving van één of meer rechten of permissies van een object of agent.

De Data Dictionary definieert semantische eenheden. Elke semantische eenheid van de Data Dictionary wordt gemapt naar één van deze entiteiten. In dit opzicht kan een semantische eenheid gezien worden als een eigenschap van een entiteit. Bv. de semantische eenheid “grootte” is een eigenschap van een object. Semantische eenheden hebben een waarde. Voor een bepaald object kan de grootte “843200004” zijn.
In sommige gevallen kunnen de semantische eenheden de vorm aannemen van een container dat een set van gerelateerde semantisch eenheden groepeert. De gegroepeerde semantische eenheden worden dan de semantische componenten van de container genoemd. Sommige van die containers kunnen gedefineerd worden als uitbreidbare containers. Deze containers laten toe om metadata in te voegen van een extern schema. Dit laat toe om PREMIS uit te breiden met metadata elementen die wat betreft de preservatie out of scope waren, maar voor de beschrijving bv. wel nuttig zijn.

Een relatie is een associatie tussen twee entiteiten. Een relatie kan op verschillende manieren geïnterpreteerd worden. Bv. de bewering “Object A heeft als formaat B” kan als een relatie tussen A en B gezien worden. Het PREMIS-model daarentegen behandelt formaat B als een eigenschap van Object A. PREMIS reserveert het concept relatie voor de associaties tussen twee of meer objecten of tussen entiteiten van verschillende types, zoals een object en een agent.

Een object heeft drie subtypes: bestand, bitstream en representatie.

- Een bestand is een genoemde en geordende sequentie van bytes dat gekend is door het operating system. Een bestand kan 0 of meer bytes hebben en heeft een bestandsformaat, toegangspermissies en file system karakteristieken zoals grootte en datum van de laatste modificatie.

- Een bitstream is data binnen een bestand dat betekenisvolle gemeenschappelijke eigenschappen hebben voor preservatie doeleinden. Een bitstream kan niet worden omgevormd naar een bestand zonder de toevoeging van een bestandsstructuur (headers, ...) en het herformateren van de bitstream naar een bepaald bestansformaat. Bv. audio data binnen een WAVE bestand, een beeld in een TIFF 6.0 bestand.

- Een representatie is een set van bestanden, met de structurele metadata, die nodig zijn voor een volledige en redelijke vertolking van een intellectuele entiteit. Bv. een artikel kan volledig in één PDF bestand zitten. Dit PDF bestand is dan de representatie van het artikel. Een ander artikel kan gerepreseteerd worden door één TIFF-bestand voor elk van de twaalf pagina`s en een XML-bestand dat de structuur beschrijft. Deze dertien bestanden vormen dan de representatie van het artikel.

Een event entiteit aggregeert metadata rond acties. De documentatie van acties die een digitaal object veranderen is essentieel voor het onderhouden van de herkomst, een sleutelelement in de authenticatie. Welke acties worden opgenomen als event is de beslissing van van de bibliotheek. In het data model kunnen objecten geassocieerd worden met events op twee manieren. Als een object gerelateerd is aan een tweede object via een event, dan wordt de event identifier opgenomen in de relaties container als de semantische component *relatedEventIdentification.* Als een object een geassocieerd event heeft zonder een relatie te hebben met een ander object, dan wordt de event identifier opgenomen in de container *linkingEventIdentifier*. Bv. een preservatie bibliotheek haalt een XML bestand

(object A) op en creëert een genormaliseerde versie ervan (object B) via een applicatie (event 1). Dit kan in *relationship* als volgt worden beschreven:

relationshipType = "derivation"
relationshipSubType = "derived from"
relatedObjectIdentification
relatedObjectIdentifierType = "local"
relatedObjectIdentifierValue = "A"
relatedObjectSequence = "not applicable"
relatedEventIdentification
relatedEventIdentifierType = "local"
relatedEventIdentifierValue = "1"
relatedEventSequence = "not applicable"

Een agent is zeker belangrijk, maar is niet de focus van de data dictionary. Deze definieert enkel een manier om een agent de identificeren en kent er een classificatie aan toe (persoon, organisatie of software). Dit is niet genoeg echter, maar wordt buiten de scope van PREMIS gezien. In het data model bestaat er een relatie tussen de entiteit Agent en de entiteit Event, maar geen van Agent naa de Object entiteit. Agenten beïnvloeden objecten alleen indirect via een event. Omdat een agent verschillende rollen kan vervullen in verschillende events, is de rol van een agent een eigenschap van de Event entiteit.

De semantische eenheden gedefinieerd in PREMIS zijn met elkaar verbonden via een aantal structurele conventies die helpen de Data Dictionary organiseren en zijn implementatie ondersteunen. Deze conventies houden het gebruik in van identifiers, die een manier zijn om relaties te leggen en het 1:1 principe om metadata te relateren aan objecten.

Instanties van objecten, gebeurtenissen (events), agenten en rechten zijn uniek identificeerbaar via een set van semantische eenheden onder de "identifier"-container. Deze semantische eenheden volgen een gelijkaardige syntaxis en structuur, ongeacht het entiteitstype:

[entity type]Identifier
[entity type]IdentifierType: domain in which the identifier is unique
[entity type]IdentifierValue: identifier string

Een voorbeeld van een object:

ObjectIdentifier
ObjectIdentifierType: NRS
ObjectIdentifierValue: http://nrs.harvard.edu/urn-3:FHCL.Loeb:sa1

Een voorbeeld van een Event:

EventIdentifier
EventIdentifierType: NRS
EventIdentifierValue: 716593

Het identifiertype “NRS” duidt aan dat de identifier uniek is binnen het domein van de Name Resolution Service die de identifiers toekent. Als de bibliotheek digitale objecten en hun metadata uitwisselt, is het nodig dat het type van de identifier wordt meegegeven.

Identifiers zijn herhaalbaar voor objecten en agenten, niet voor rechten en events. Objecten en agenten kunnen verschillende identiteiten hebben in een globale omgeving en tussen verschillende systemen. Daarom is het noodzakelijk dat aan deze entiteiten verschillende identifiers kunnen worden toegewezen. Rechten en events hebben een context die is gelimiteerd tot de bibliotheek en daarvoor zijn hiervoor geen meerdere identifiers nodig.

4.4 Conceptuele modellen

4.4.1 FRBR

FRBR staat voor “Functional Requirements for Bibliographic Records”. Het is een conceptueel model voor gebruik in de bibliografische wereld dat is neergeschreven in een rapport dat verschenen is in 1998. Het rapport is opgesteld door IFLA (International Federation of Library Associations and Institutions). Het doel van het rapport is een duidelijk gedefinieerd, gestructureerd raamwerk dat data van bibliografische records relateert aan de noden van de gebruikers van deze records. De gebruiker staat centraal in dit raamwerk. In deze optiek zijn er twee belangrijke concepten in het FRBR raamwerk: gebruikerstaken en bibliografische entiteiten.

Het FRBR model definieert de volgende taken die de rest van het model bepalen: Find, Identify, Select en Obtain die als volgt zijn gedefinieerd:

- Find: Om entiteiten te vinden die aan bepaalde criteria van de gebruiker voldoen. Deze voorwaarden zijn gebaseerd op de attributen of relaties van die entiteit.
- Identity: Om entiteiten te identificeren. Bv. om te bevestigen dat de beschreven entiteit overeenkomt met de gezochte entiteit of om een onderscheid te kunnen maken tussen entiteiten met zeer gelijke eigenschappen
- Select: Om entiteiten te selecteren die aan de gebruiker zijn gehoefte voldoen. Bv. om een entiteit te kiezen die aan bepaalde gebruikersverwachtingen voldoet zoals de inhoud, het formaat, ...
- Obtain: Om toegang tot beschreven entiteiten te verkrijgen.

De bibliografische entiteiten kunnen ingedeeld worden in groepen. De best gekende features van het model zijn de groep 1 entiteiten. Deze representeren de producten die het resultaat zijn van een intellectuele of artistieke inspanning en beschreven zijn in bibliografische records. Deze entiteiten zijn conceptueel. Dit wil zeggen dat ze niet speciaal zijn voorzien om records opgeslagen in een databank te representeren. De groep 1 entiteiten zijn:

- Werk: een intellectuele of artistieke creatie
- Expressie: de intellectuele of artistieke realisatie van een Work.
- Manifestatie: de fysische belichaming van een Expression of Work
- Item: een versie van een Manifestation

Een Werk is een abstract concept, het idee achter iets, voordat het in één of andere vorm is gefixeerd. Een Expressie is deze fixatie, die het idee omzet in een representatie zoals woorden of muzieknoten. Deze fixatie is nog steeds conceptueel en nog niet fysiek. Een

Werk kan verschillende Expressies hebben, bv. in verschillende talen. Een Manifestatie is een set van fysieke zaken die een Expressie of Werk bevatten. Een Manifestatie kan verschillende Expressies bevatten, zoals het geval is bij een CD waarop verschillende liedjes staan. Deze liedjes zijn dan een Expressie van een individueel Werk. Een Item is dan een individuele copy van een manifestatie. Dit kan een fysiek iets zijn, maar ook een copy van een file.
De groep 2 entiteiten zijn gedefinieerd als diegene die verantwoordelijk zijn voor de intellectuele of artistieke inhoud, de fysieke productie en distributie of het beheer van de entiteiten van groep 1. Het FRBR rapport bevat slechts twee groep 2 entiteiten: Person en Corporate Body, alhoewel Family vaak als derde groep 2 entiteit wordt opgegeven. Elk van de groep 2 entiteiten kunnen als verantwoordelijk gesteld worden voor een groep 1 entiteit. Bv. de auteur is de creator van een Work, de vertaler realiseert een Expression, de uitgever een Manifestation en een bibliotheek bezit dan een Item ervan.

Groep 3 entiteiten worden beschreven als de onderwerpen van de Werken. Elke groep 1 of groep 2 entiteit valt onder deze categorie, maar ook de bijkomende entiteiten Concept, Object, Event en Place.

Het FRBR rapport geeft dus een conceptueel model en geen concreet data model. Het model is daarom onderworpen aan beslissingen op implementatiegebied. Verschillende implementaties vertegenwoordigen verschillende functionaliteiten. De functionaliteiten die mogelijk zijn met systemen die het FRBR principe implementeren. Bibliotheek catalogen die gebruik maken van de FRBR principes kunnen de zoekresultaten gemakkelijker groeperen. Er kan een lijst met alle werken van een bepaalde creator kan worden weergegeven. Daarnaast kan men de verschillende expressies van een werk geven, gegroepeerd volgens formaat, taal, uitvoerder, director of volgens om het even welk attribuut.

Het FRBR model is bruikbaar voor alle bibliotheken, maar OCLC studies hebben aangetoond dat niet alle werken voordeel halen uit de FRBR principes. Aan records beschreven volgens het FRBR model is een zekere kost verbonden om deze modellen te implementeren. Werken uit het veld van de literatuur en muziek zullen het meest bruikbaar zijn, daar van deze werken verschillende versies in de tijd bestaan, maar ook van uitvoerder, etc.

4.4.2 CIDOC-CRM

Het CIDOC objectgeoriënteerd conceptueel referentiemodel werd ontwikkeld door de ICOM\CIDOC Documentation Standards Group. De standaard representeert een ontologie

voor culturele erfgoedinformatie. Het CIDOC Conceptueel Referentie Model (CRM) levert de definities en een formele structuur om expliciete en impliciete concepten en relaties die gebruikt worden bij de documentatie van cultureel erfgoed te beschrijven. CIDOC CRM probeert de kennis over informatie met betrekking tot cultureel ergoed te promoten door een gemeenschappelijk en uitbreidbaar semantisch raamwerk op te zetten waarnaar alle informatie over cultureel erfgoed kan gemapt worden. Het is de bedoeling om een gemeenschappelijke taal te ontwikkelen voor domeinexperten en developers om de requirements van informatiesystemen te formuleren en om te dienen als een gids voor het conceptueel modeleren. Op deze manier wil men de semantiek aanleveren die nodig is om te bemiddelen tussen de verschillende informatiebronnen over cultureel erfgoed zoals die gepubliceerd zijn door musea, bibliotheken en archieven.

Een duidelijke visie op de ontologie is noodzakelijk. Ze kan de nodige antwoorden aanreiken voor vragen omtrent wat wel en niet door de ontologie wordt gedragen. Hier wordt een onderscheid gemaakt tussen de theoretische en de praktische scope.

De theoretische scope van CIDOC CRM moet gezien worden als het domein dat CIDOC CRM wil behandelen indien er genoeg tijd en middelen aanwezig zijn. De praktische scope is een subset van de theoretische scope en is gedefinieerd als het domein dat op dit moment door CIDOC CRM wordt gecovered. Dat wil zeggen dat er mappings worden voorzien die de vertaling verzorgen van en naar de brondocumenten. Deze scope kan dus veranderen naargelang er andere standaarden relevanter worden.

De theoretische scope van CIDOC CRM wordt gedefinieerd als de nodige informatie voor de wetenschappelijke beschrijving van collecties aan cultureel erfgoed samen met de nodige vertalingen om informatie te kunnen uitwisselen tussen heterogene informatiebronnen. Met cultureel erfgoed worden alle types van materiaal bedoeld die verzameld en tentoongesteld worden door de musea en aanverwante instituten. Dit betreffen dus collecties, sites en monumenten die te maken hebben met de geschiedenis, etnografie, archeologie, historische monumenten en collecties kunstwerken. Voor de definitie van cultureel erfgoed wordt verwezen naar de definitie die is opgesteld door ICOM. Als doel stelt CIDOC CRM voorop dat de kwaliteit van de beschreven informatie goed genoeg moet zijn voor academisch onderzoek en niet voor het occasioneel browsen van de informatie. CIDOC CRM richt zich dus vooral op de museumprofessionals en de onderzoekers.

CIDOC CRM richt zich ook voornamelijk op de beschrijving van de contextuele informatie. Dit houdt voornamelijk de historische, geografische en theoretische achtergrond in van de

tentoongestelde items, waardoor hun waarde en betekenis vergroten. CIDOC CRM houdt zich dus bezig met de beschrijvende metadata. Wat buiten de scope van CIDOC CRM valt, is de informatie voor de administratie en het beheer.

De praktische scope van CIDOC CRM is dus een subset van de theoretische. De elementen van de volgende datastructuren zijn reeds opgenomen in CIDOC CRM. Deze worden geverifieerd door de mappings die worden bijgevoegd bij de ondersteunende documentatie. De volgende lijst geeft hun status weer:

Volledig:

- Dublin Core
- Art Museum Image Consortium (AMICO)
- Encoded Archival Description (EAD)
- MDA SPECTRUM
- Natural History Museum (London) John Clayton Herbarium Data Dictionary
- National Museum of Denmark
- International Federation of Library Associations and Institutions (IFLA) Functional Requirements for Bibliographic Records (FRBR)
- OPENGIS
- Association of American Museums Nazi-era Provenance Standard
- MPEG-7
- Research Libraries Group (RLG) Cultural Materials Initiative DTD

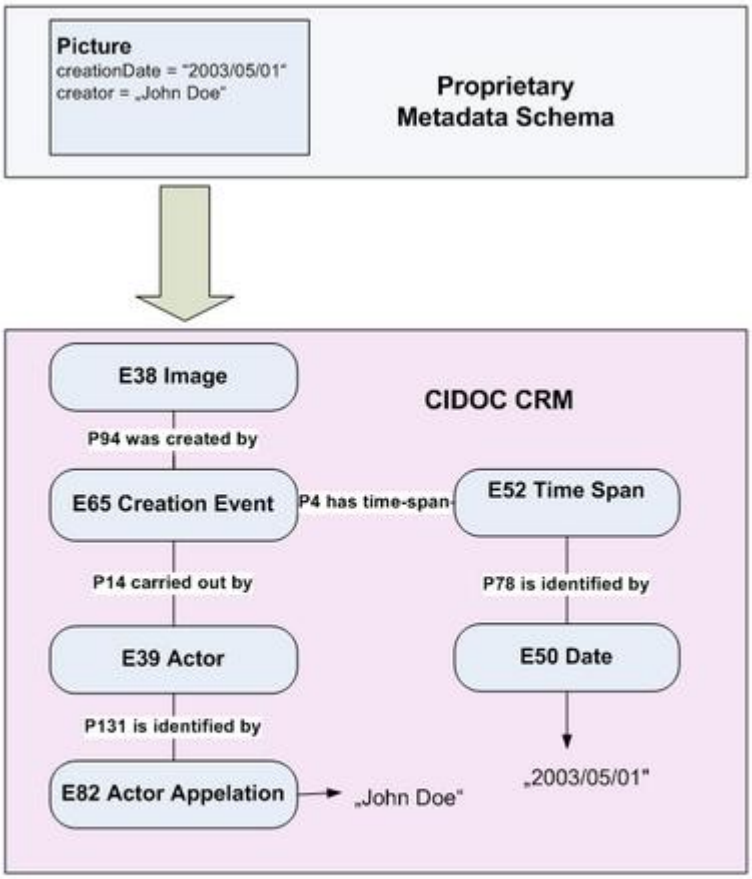
In Progress:

- Consortium for the Computer Interchange of Museum Information (CIMI) Z39.50 profile
- Council for Prevention of Art Theft Object ID
- The International Committee for Documentation of the International Council of Museums (CIDOC) The International Core Data Structures for Archaeological and Architectural Heritage
- Core Data Index to Historic Buildings and Monuments of the Architectural Heritage
- English Heritage MIDAS – A Manual and Data Standard for Monument Inventories
- English Heritage SMR 97
- Hellenic Ministry of Culture POLEMON Data Dictionary

Gewenst:

- FENSCORE
- Sydney University TimeMapper
- Data Service Standards in Archaeology
- Digital Library Metadata
- International Council on Archives (ICA) ISAD(G) – International Standard Archival Description (General)
- Visual Resources Association Core Categories – VRA Core
- Machine Readable Cataloguing – MARC
- CIMI SGML DTD
- Getty Categories for the Description of Works of Art – CDWA
- RSLP Collection Description
- MODES OBJECT FORMAT

Het raamwerk voorziet 84 klassen en 141 properties. Klassen worden doorgaans met de notatie Enn weergegeven, properties met Pnn en de gegeven naam van de klasse of property tussen haakjes. Het grootste voordeel van CIDOC CRM is de genericiteit. Hierdoor kan vrijwel elke standaard of proprietair metadataschema in de culturele erfgoedsector gemapt worden naar CIDOC CRM. Een ander voordeel van deze standaard is de granulariteit die zeer specifieke queries toelaat.
Het grootste nadeel van de CRM is zijn complexiteit. Algemeen wordt aangenomen dat deze standaard niet bedoeld is om rechtstreeks aan de eindgebruiker getoond te worden wegens de overvloed aan informatie. De meeste bestaande applicaties die CIDOC CRM gebruiken, zoals I-MASS en SCULPTEUR, tonen ofwel een vereenvoudigde versie van het schema ofwel begeleiden ze de eindgebruiker met een aantal vragen naar de juiste informatie. Typisch zullen bepaalde properties in een metadataschema niet mappen naar één enkele property maar naar een ketting van klassen en properties die hen verbinden. Als een metadatarecord die een object beschrijft bijvoorbeed de velden “CreationDate” en “Creator” bevat, wordt dit naar CIDOC CRM gemapt als “P94(was created by) – E65(CreationEvent) – P14(carried out by) – E39(Actor) – P131(is identified by) – E82(Actor Appellation)” en “P94(was created by) – E65(CreationEvent) – P4(has time-span) – E52(Time-span) – P78(is identified by) – E50(Date)”. Deze mapping wordt in de onderstaande tekening schematisch weergegeven.



4.4.3 ABC

Naast SPECTRUM en ISAD(G) is ook het ABC-model het vermelden waard. Dit model is het resultaat van “The Harmony Project” en is ontworpen om in een gemeenschappelijk conceptueel model te voorzien dat de interoperabiliteit tussen verschillende metadata ontologieën van verschillende domeinen moet vergemakkelijken. Hij is niet bedoeld als een metadatawoordenschat maar als een model dat als basis kan dienen bij het ontwerpen van specifieke ontologieën.

De kernbedoeling bij het ontwerpen van dit model was het voorzien van de mogelijkheid om het hele traject van een object weer te geven (naast de traditionele metadata). Dit maakt het mogelijk de creatie, de evolutie en de overgangen die objecten meemaken te beschrijven. Zo kan worden beschreven waar en wanneer het interview is afgenomen, wie het afnam, transfers naar andere media, enz. Op deze manier kan de gehele levenscyclus van een

object worden opgevraagd en dankzij de bidirectionele relaties kan ook informatie opgevraagd worden over elk object dat verbonden is met een object gedurende zijn levenscyclus.

Het ABC-model voorziet ook in een hiërarchie voor objecten en voor eigenschappen. Zo kunnen objecten worden georganiseerd in een hiërarchie waarbij elke subklasse extra informatie levert over het object. Deze hiërarchie vergemakkelijkt de interoperabiliteit tussen verschillende standaarden dankzij “partial understanding”. Beschikt een bepaald object niet over een tegenhanger in de andere standaard dan kan men op zoek gaan naar een tegenhanger op een hoger niveau.

De hiërarchie voor eigenschappen laat eveneens toe deze te verfijnen met zogenaamde subeigenschappen. Hierdoor kan men niet alleen een gewenst informatieniveau bepalen (internen krijgen toegang tot alle informatie, externen enkel tot een bepaald niveau) maar ook de zoeknauwkeurigheid.

De hiërarchieën, levenscycli en bidirectionele relaties laten verder ook toe om van eenzelfde object/item verschillende representaties ter beschikking te hebben.

4.4.4 GAMA

GAMA, Gateway to Archives of Media Art, is een interdisciplinair project dat startte op 01/11/2007. Aan het project nemen 19 organisaties deel uit de Europse cultuursector, kunstensector en technologiesector. Het doel van het project is om een centrale, online protal te bouwen die toegang verleent tot verschillende Europese mediakunstencollecties voor het geïnteresseerde publiek, curatoren, artiesten, academici en onderzoekers. De uitkomst van het project was onder andere een ontologie die zich richt op het beschrijven van mediakunsten. Dit project wordt voor 1.2 miljoen Euro gesteund door de Europese Commissie via het *e*contentplus programma.

Media kunst, die kunstwerken creëert met behulp van nieuwe media technologieën, is een van de meest populaire eigentijdse kunstgenres. Media kunst onderzoekt de spanning tussen cultuur en technologie en de ambigue rol van nieuwe mediatechnologieën. Tegelijkertijd brengt de media kunst ook die nieuwe technologieën onder de publieke aandacht. In dit opzicht hebben de mediakunst archieven een belangrijke rol. Het gebruik en hergebruik van de digitaal materiaal uit de mediakunst archieven staat in schril contrast tot

de relevantie van mediakunst in de eigentijdse cultuur. Deze discrepantie wil het GAMA project wegwerken.

Het consortium van het GAMA project beslaat de meerderheid van de belangrijkste, Europese content leveranciers wat betreft mediakunst. De media aangeboden door de partners omsluit zo`n 55% van alle mediakunstwerken die online toegankelijk zijn door Europese culturele archieven en verdelers. Het doel is dus het opzetten van een centraal platform dat toegang verleent tot die mediakunst archieven. Dit platform moet georiënteerd zijn op de gebruiker en moet verscillende talen ondersteunen. Dit platform moet het gebruik, het hergebruik en de zichtbaarheid van de mediakunst en hun aangereikte media vergroten. Deze gateway moet evolueren tot het centraal zoekportaal voor Europese mediakunst.

Het portaal moet de toegang tot de verschillende archieven vergemakkelijken en verbeteren ongeacht de structuur van de archieven, de gebruikte metadata, de gebruikte taal, de verschillende gebruikte digitale formaten en hun focus. Om aan deze eisen te voldoen ontwikkelde GAMA een nieuwe ontologie speciaal ontworpen voor beschrijven van mediakunstwerken. Deze wordt kort besproken in de volgende paragraaf.

Het GAMA metadata schema is beschreven in RDF. Klassen, eigenschappen en datatypes zijn de bouwblokken van het GAMA metadata schema. Allen behoren ze tot hetzelfde namespace: http://gama-gateway.eu/schema/ of kortweg gama. Het schema staat hieronder weergegeven. Het kent elf verschillende klassen. Deze klassen kunnen opgedeeld worden in twee groepen: entiteiten en enumeraties.

De entiteiten bestaan uit de volgende klassen:

gama:Work (representeert kunstwerken, events, en andere bronnen)

gama:Person (representeert een person, instituut of collectief)

gama:Manifestation (representeert fysieke representaties van werken)

gama:Archive (representeert een archief)

gama:Collection (representeert collecties van werken)

De enumeraties zijn klassen met gefixeerde instanties en bestaan uit de volgende klassen:

gama:WorkType (lijst van types van werken)

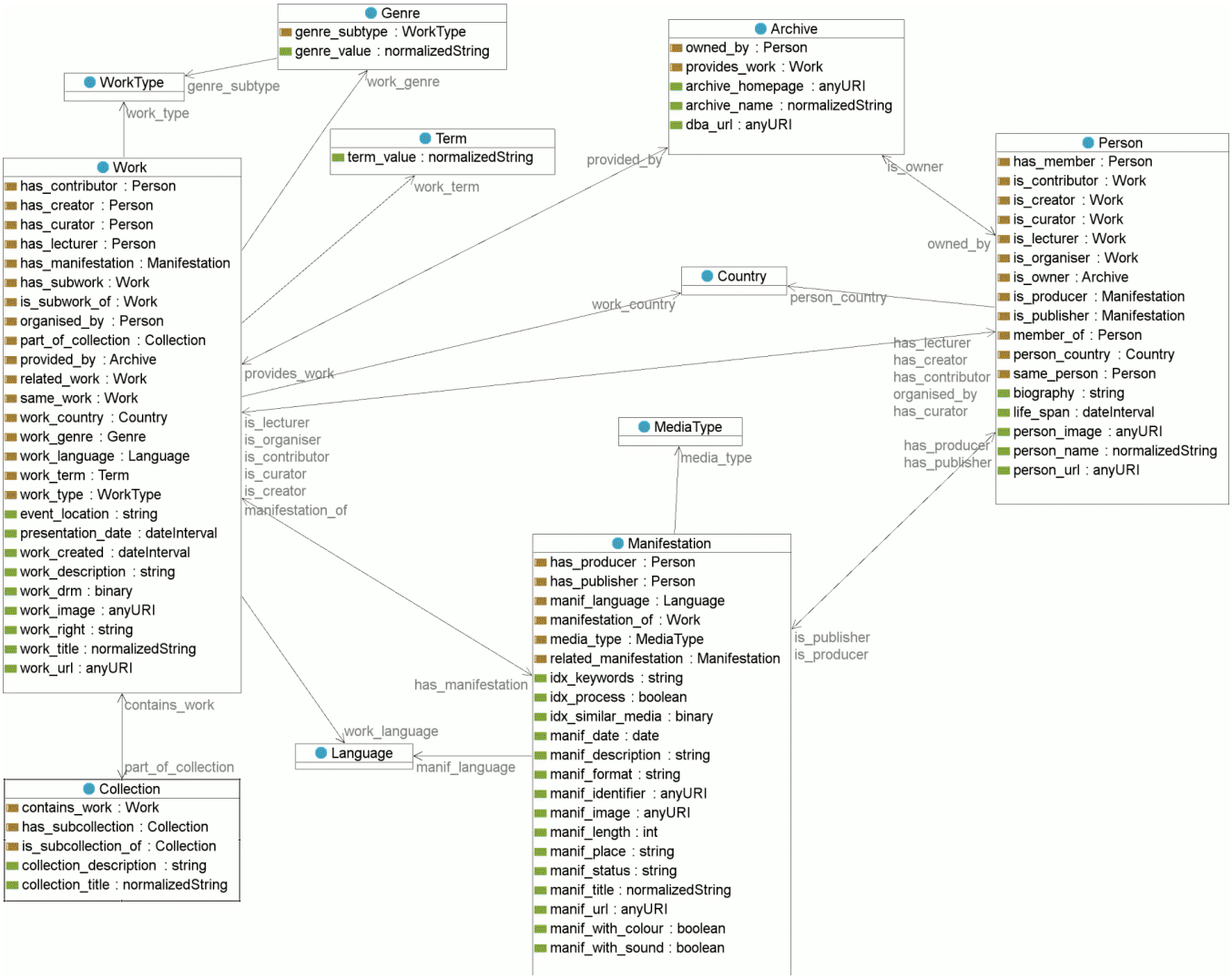
gama:MediaType (lijst van mediatypes)

gama:Genre (twee-laagse hiërarchie van genres)

gama:Term (lijst van veelgebruikte termen die gerelateerd zijn aan werken)

gama:Country (lijst van landen)

gama:Language (lijst van talen)



Metadata schema GAMA

4.5 Thesauri

4.5.1 FRAR

In bibliotheken, musea of archieven catalogen zijn steeds een set georganiseerde data dat de informatie beschrijft die wordt beheert door het instituut. Om verschillende werken van een persoon te groeperen of verschillende edities van een bepaald werk te groeperen, is er nood aan gecontroleerde toegangspunten voor auteurs en titels. Veel namen van auteurs hebben verschillende schrijfwijzen en hetzelfde geldt voor de titels. Deze toegangspunten moeten dus alle schrijfwijzen aanleveren, zowel de geauthoriseerde als de variabele, van de namen van de auteurs en de titels. Dus authoriteitscontrole houdt zowel het beheer in van de geauthoriseerde schrijfwijzen als het identificeren van entiteiten die worden voorgesteld door deze toegangspunten. Hieruit haalt de eindgebruiker zeker voordeel doordat hij elke vorm kan gebruiken van een auteursnaam of titel om informatie hieromtrend te vinden.

Rond dit doel werd FRANAR opgericht. FRANAR, de werkingsgroep rond Functional Requirements and Numbering of Authority Records, werd opgericht in april 1999. Deze groep had verschillende doelen. Twee daarvan was het onderzoek naar de benodigdheden om te kunnen spreken over authority records en het onderzoek naar de haalbaarheid van een internationaal gestandardiseerde nummering voor authority data, ISADN, International Standard Authority Data Number.

Authoriteitsrecords zijn geaggregeerde informatie over een bepaalde instantie of entiteit waarvan de naam wordt gbruikt als toegangspunt tot de bibliografische records ervan. Uit deze werkingsgroep is reeds een eerste oplevering gebeurd: FRAR, Functional Requirements for Authority Data. Het is een document dat een conceptueel model beschrijft voor het aanleveren van een analytisch raamwerk voor de analyse van de functionele benodigdheden voor authoriteitsdata die nodig zijn voor het ondersteunen van authoriteitscontrole en voor het internationaal delen van authoriteitsdata.

Meer specifiek biedt het document een conceptueel model dat is ontwikkeld om:

- Een referentiekader aan te leveren om data afkomstig van authority records te relateren naar de noden van de gebruiker van deze records.
- Te assisteren in het internationaal delen van de authority data zowel binnen als buiten het bibliotheekwezen.

Het model focust op de data, ongeacht hoe deze is verpakt in b.v. records. Het is in feite een uitbreiding van de FRBR standaard. De methodologie die werd gebruikt bij het opstellen van het conceptueel model is dezelfde als de gevolgde methodologie bij FRBR:

- Het eerste dat moet gebeuren is de identificatie van de objecten waarin de eindgebruiker is geïnteresseerd. Elk van deze objecten, of entiteiten, dienen als een

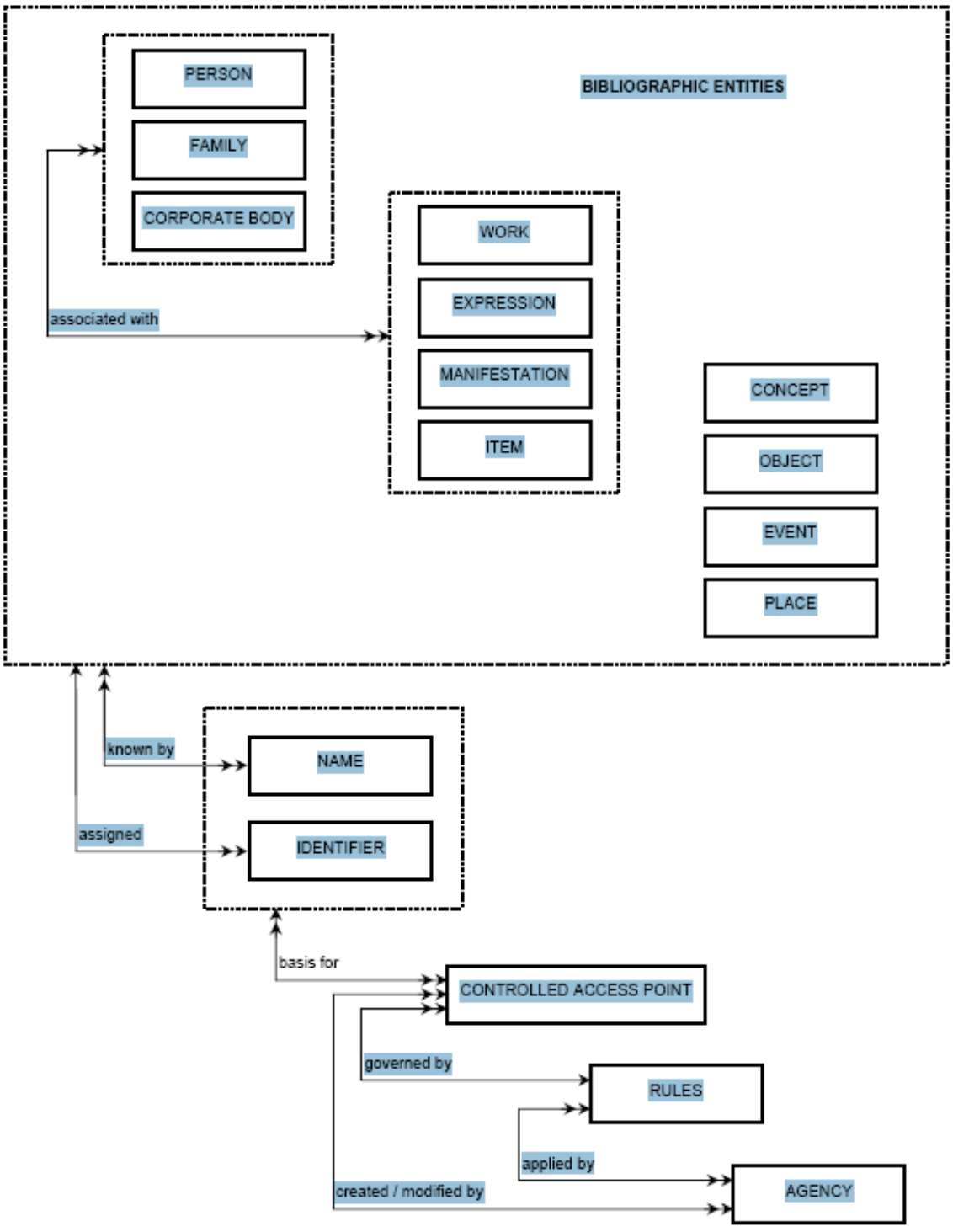
soort toegangspunt tot een cluster van data. Het bekomen model kan ook relaties leggen tussen verschillende types van entiteiten.

- Wanneer het identificeren van de entiteiten en de relaties tussen de verschillende entiteiten is voltooid, is de volgende stap het identificeren van de primaire eigenschappen of attributen van elke entiteit.

Entiteiten uit de bibliografische wereld, zoals deze bij FRBR, zijn gekend door hun namen en eventueel identifiers. Bij authoriteitscontrole worden deze namen en identifiers gebruikt als toegangspunt tot de informatie. De entiteiten waarop authoriteitsdata zich focust zijn de entiteiten die werden geïdentificeerd bij FRBR: person, corporate body, work, expression, manifestation, item, concept, object, event en place (en eventueel family). Deze zijn terug te vinden op het bovenste gedeelte van het onderstaande diagramma. Het onderste gedeelte vertoont de namen van deze entiteiten, de identifiers ervan en de gecontroleerde toegangspunten, die zijn gebaseerd op deze namen en identifiers die werden geregistreerd in authority files. Ook de regels die werden gehanteerd bij het opstellen van de cataloog spelen een rol in dit diagramma. Zo kan een auteur in een bepaalde cataloog de fysische persoon zijn met al zijn pseudoniemen, terwijl in een andere cataloog elk pseudoniem als een andere auteur kan worden aanzien. De regels zijn dus ook belangrijk. De tekening verduidelijkt de relaties die bestaan tussen de naam (en identifier) en de bibliografische entiteit (person, ...). Een specifieke instantie van elk van deze entiteiten is gekend via één of meerdere namen en omgekeerd kan elke naam worden gerelateerd aan één of meerdere instanties van die entiteiten. Gelijkaardig kan een specifieke instantie van een entiteit worden gerelateerd aan één of meerdere identifiers, maar een identifier kan slechts gerelateerd worden aan één specifieke instantie van een entiteit. Het diagramma houdt ook rekening met relaties die kunnen bestaan tussen personen, organisaties, families enerzijds en werken, manifestaties, expressies en items anderzijds.

Het onderste deel van het diagramma toont ook de associaties die bestaan tussen de namen (en identifiers) van de entiteiten en de gecontroleerde toegangspunten van de entiteiten en de associaties tussen entiteiten en de regels voor die entiteiten. Een toegangspunt kan b.v. gebaseerd zijn op de combinatie van twee namen en/of identifiers, zoals het geval kan zijn bij werken die worden geïdentificeerd door hun naam en titel. Deze toegangspunten worden bepaald door regels en die regels kunnen toegepast worden door verschillende agentschappen. Een toegangspunt kan dus worden gecreëerd of veranderd door een agentschap.

Er kan nog een ander type relatie bestaan: deze tussen een instantie van een bepaald type entiteit en een instantie van een ander type van entiteit. Dit kan b.v. een relatie zijn van een persoon tot een organisatie. Deze relaties werden echter niet vertoond op het diagramma.



Het model levert entiteitsdefinities die voornamelijk worden gehaald uit twee standaarden: FRBR, Functional Requirements for Bibliographic Records, en GARR, Guidelines for Authority Records and References. De entiteiten die worden aangehaald in dit model zijn: persoon, familie, organisatie, werk, expressie, manifestatie, item, concept, object, event, plaats, naam, identifier, gecontroleerd toegangspunt, regels en agentschap. Elk van deze entiteiten bezit een aantal attributen die werden vooral gehaald uit FRBR, GARR en ISAAR(CPF). Verder specifieert het conceptueel model ook alle mogelijke relaties die tussen de entiteiten kunne bestaan. Voor een verdere uitdieping van de definities van de entiteiten, hun attributen en relaties wordt verwezen naar de specificaties van FRAR.

In praktijk kan het proces voor het catalogiseren worden opgesplitst in 3 delen:

- Bibliografische beschrijvingen generen: De catalogiseur creëert bibliografische beschrijvingen voor de bronnen die aanwezig zijn in de bibliotheek. De regels die hierbij worden toegepast zijn opgesteld uit de bronnen waarvan de data wordt afgeleid, de volgorde en de vorm van de individuele data elementen.

- Toegangspunten formuleren: Hierbij worden de toegangspunten gecreëerd van zowel de geauthoriseerde vormen van de naam die een auteur, titel, onderwerp en dergelijke representeert als de variabele vormen van de geauthoriseerde vorm.

- Toegangspunten registreren: De catalogiseur registreert de gecontroleerde toegangspunten in authority bestanden. Het registreren van een nieuw toegangspunt kan wel tot gevolg hebben dat reeds bestaande toegangspunten moeten worden gereviseerd.

4.5.2 LCSH

De Libarary of Congress Subject Headings behelst een thesaurus van indexeringen (subject headings) die wordt onderhouden door de Library of Congress uit de Verenigde Staten. Bibliografische records vormen het toepassingsgebied van de indexeringen. Deze indexeringen zijn van toepassing op alle items van een bibliografisch record. Op deze manier wordt de toegang tot items die handelen over een bepaald onderwerp in een cataloog vergemakkelijkt.

LCSH wordt reeds veel gebruikt in bibliotheken. Hierdoor wordt het zoeken naar items uit een bibliotheek die dezelfde zoekstrategie en de LCSH thesaurus gebruikt geuniformiseerd. Ondanks de uitgebreide scope van LCSH en zijn groot gebruik bestaan er bepaalde

bibliotheken die nood hebben aan een andere indexering. Zo heeft de National Library of Medicine van de Verenigde Staten een eigen thesaurus ontwikkeld, Medical Subject Headings (MeSH). Vele universiteiten gebruiken de twee indexeringen voor hun items. In Canada heeft de National Library of Canada samengewerkt met een aantal afgevaardigden van de LCSH om een aanvullende set van Canadese Subject Headings te ontwikkelen, CSH, die gebruikt wordt voor typisch Canadese onderwerpen.

De Subject Headings worden gepubliceerd in vijf grote rode volumes of kunnen online geraadpleegd worden op de site van de Library of Congress. Deze lijst met topics wordt wekelijks aangepast.

Principes

De subject headings zorgen voor een toegangspunt met een gestandardiseerde vorm wat betreft de termen, de namen en uniforme titels die het onderwerp of genre van een werk reflecteren. Mensen met een interesse in een bepaald onderwerp zonder weet van een bepaalde titel of auteur zouden op deze manier in staat moeten zijn om informatie over gerelateerde onderwerpen terug te vinden. Dus wanneer een werk handelt over een persoon, organisatie, plaats of een ander werk, dan kunnen de namen van de personen, organisaties en plaatsen alsook de uniforme titels als subject heading worden gebruikt.

Er wordt wel een onderscheid gemaakt tussen catalogiseren en indexeren. LCSH is opgesteld om het hoofdonderwerp van een bepaald werk aan te duiden. In het algemeen moet 20% van het werk over het onderwerp gaan, indien de uniforme naam van het onderwerp als subject heading voor dat werk kan gebruikt worden. In dit opzicht is het toewijzen van subject headings niet hetzelfde als een gedetaileerde indexering. Deze indexeringen behandelen soms een onderwerp dat slechts éénmaal wordt vernoemd in een bepaald werk. Het is dus belangrijk te postuleren dat de subject headings die worden toegewezen aan een bepaald werk een beknopte, gestandardiseerde samenvatting geven over dat werk. Men mag hier niet te gedetaileerd in gaan.

Niettegenstaande het feit dat men niet te gedetaileerd mag gaan in het toewijzen van subject headings, is het toch zaak om zo specifiek mogelijk te zijn. Elk onderwerp kan worden onderverdeeld in verdere subcategorieën. Over het gebruik van de categorieën of hun subcategorieën bestaan enkele regels die worden verduidelijkt met een voorbeeld:

- Als 3 of minder subcategorieën van een bepaald onderwerp worden besproken:

- Als de subcategorieën tesamen het hele onderwerp beschrijven, dan wordt de naam van het onderwerp toegezen als subject heading.

  Bv. Een boek over gewervelde en ongewervelde dieren. Dieren zijn ofwel gewerveld of ongewerveld dus wordt als subject heading dieren toegewezen en niet de termen gewervelde dieren en ongewervelde dieren.

- Als de subcategorieën tesamen slechts een deel van een bepaald onderwerp beschrijven, dan worden de subcategorieën gebruikt als subject headings voor het werk.

  Bv. Een boek over muizen en ratten. Beide zijn knaagdieren, maar er bestaan buiten hen nog veel andere soorten knaagdieren. Dus als subject headings worden hier twee termen meegegeven nl. muizen en ratten in plaats van knaagdieren.

- Als 3 of meer subcategorieën van een bepaald onderwerp worden besproken, dan wordt als subject heading de naam van het onderwerp meegegeven.

  Bv. Een boek over muizen, olifanten, beren en herten. Dan wordt als subject heading de term zoogdieren meegegeven.

Ook omtrend het aantal subject headings die worden toegewezen aan een bepaald werk bestaan een aantal richtlijnen.

- Veel werken behandelen meer dan één onderwerp, een groep van onderwerpen. Elk onderwerp moet dan als subject heading worden toegezen aan het werk.

- Soms worden verschillende facetten van hetzelfde onderwerp besproken in een werk en dan heb je ook meerdere subject headings nodig om het werk goed te catelogiseren.

  Bv. Een werk rond een bepaald economisch probleem op een bepaalde plaats. Dan kan je starten met een subjet heading die het probleem aanduidt. Als het probleem een lokale subdivisie heeft gerelateerd aan die plaats, kan dat lokale probleem gekozen worden als subject heading om het probleem weer te geven. Een tweede subject heading kan de plaats zijn, waarvan de subdivisie “Economic conditions” is.

- Een werk kan ook een onderwerp behandelen dat zich over meerdere niveau laat beschrijven. Bv. een algemene discussie over een concept met een case study die

het concept illustreert in een bepaalde context. De headings moeten de verschillende niveaus reflecteren als minstens 20% van het werk is gewijd aan elk niveau.

Bv. Women $z Nicarague; Women

In het algemeen wordt aangenomen dat een boek bv. adequaat kan beschreven worden door middel van 6 subject headings. Ook moet vermeden worden om meer dan 10 subject headings toe te wijzen aan een bepaald werk. De volgorde van de subject headings moet ook de belangrijkheid van de subject headings weergeven.

Vorm

Dit stukje gaat over de vorm van de subject headings, de syntaxis. Er wordt een onderscheid gemaakt tussen de hoofdheading en een subdivisie ervan.

- Hoofdheading:
    - Onderwerp - Topic (MARC tag 650): Dit representeert een concreet object, dier,..., een categorie van mensen, dieren of objecten, een abstrakter concept, geloof, proces of fenomeen, een instituut, ...

      Een topical heading kan een enkele term zijn of een zin:

        - Enkele term: waarschijnlijk de meest gangbare vorm.

      Bv. Women

      Savannas

      Housing

        - Een zin: Er bestaan verschillende methodes om een zin te construeren.
            - Directe volgorde:

          Bv. Housing policy

          Foreign exchange administration

            - Omgekeerde volgorde: omgekeerde headings worden gescheiden door een komma. De meer significante term wordt eerst gegeven, gevolgd door een bepaling.

          Bv. Authors, Mexican

          Farms, small

- Term plus een bepalende term: de bepalende term wordt tussen ronde haakjes meegegeven en moet de context aanduiden van het onderwerp.

Bv. Stress (Physiology)

Stress (Psychology)

- Een zin over een topic mag voorzetsels hebben.

Bv. Violence in motion pictures

- Een zin mag gerelateerde termen bevatten die verbonden zijn met het woord “and”.

Bv. Banks and banking

Cities and Towns

- Soms bestaat een zin uit termen die gerelateerd worden beschouwd. De zin bestaat dan uit de twee gerelateerde termen plus “etc.”

Bv. Comic books, strips, etc.

- Een combinatie van de bovenvermelde structuren.

- Naam – van persoon, organisatie of conferentie (MARC tag 600, 610 of 611): De vorm van alle namen moet hetzelfde zijn ongeacht de functie van het object dat een naam krijgt toegewezen. (bv. author, responsible body of subject)
- Uniforme titel (MARC tag 630): Hiervoor geldt dezelfde regel als voor namen.
- Een geografische plaats (MARC tag 651): Hier bestaan twee categorieën:
  - Plaatsen die een jurisdictieve status hebben of hadden zoals landen, steden, provincies. Zo`n plaatsen hebben een overheid die als “corporate authos” kan beschouwd worden.

    Bv. Argentinië

    Buenos Aires (Argentinië)
  - Plaatsen zonder jurisdictieve status.

    Bv. Olympus, Mount (Greece)

    Atlantic coast (Nicaragua)

- Subdivisies: subject heading strings

Een subject heading kan bestaan uit een string, met een heading en één of meer subdivisies. Deze subdivisies worden met een bepaalde string gespecifieerd.

Bv. Farms, Small $z Colombia

Camus, albert, $d 1913-1960 $v Congresses

Women $z Italy $x Social conditions

Enkele beperkingen

De lijst van LC subject headings is beperkt. Het is gebaseerd op een literaire garantie wat wil zeggen dat LCSH is gebaseerd op headings die werkelijk zijn gebruikt om de onderwerpen te beschrijven van de werken die zijn gecatalogiseerd in het LC.

De keuze en de vorm van een heading is niet noodzakelijk up-to-date. Wekelijks wordt de lijst geüpdated. Bv. Man is nu verouderd en vervangen door Human being.

De vorm van de subject headings kan nogal eens veranderen. Terwijl vroeger vaak de omgekeerde volgorde werd gehanteerd om subject headings te maken die bestaan uit zinnen, wordt tegenwoording vaak de directe volgorde gebruikt. BV. Societies, Primitive is nu verouderd en veranderd naar Primitive societies.

4.5.3 GETTY Thesauri

De woordenschat databases van het Getty instituut worden geproduceerd en onderhouden door de Getty Vocabulary Program. Ze volgen de ISO en NISO standaard voor de constructie van thesauri. Ze bevatten termen, namen en andere informatie omtrend mensen, plaatsen en concepten die gerelateerd zijn aan kunst, architectuur en materiële cultuur.

De Getty thesauri kunnen op drie verschillende manieren worden gebruikt:

- Op data invoer niveau om zaken te beschrijven.
- Als kennisbank die informatie levert aan onderzoekers.
- Als search assistants om de toegang van de eindgebruiker tot online bronnen te vergemakkelijken.

De drie voornaamste thesauri zijn:

De Art & Architecture Thesaurus (AAT): De AAT is een gestructureerde woordenschat opgebouwd rond 34000 concepten, waaronder 131000 termen, beschrijvingen,

bibliografische citaten en andere informatie over beeldende kunsten, architectuur, decorative kunsten, archivistische zaken en materiële cultuur.

De Union List of Artist Names (ULAN): De ULAN bevat zo`n 120000 records, waaronder 293000 namen en biografische en bibliografische informatie omtrend artiesten en architecten, waaronder ook tal van variaties op de namen en pseudoniemen.

De Getty Thesaurus of Geografic Names (TGN): De TGN bevat zo`n 912000 records, waaronder 1.1 miljoen namen, plaats types, coördinaten en beschrijvende nota`s gericht op plaatsen die belangrijk zijn voor de studie van kunst en architectuur.

Deze drie thesauri worden hier kort uitvoeriger besproken:

AAT

Zoals reeds eerder vernoemd bestaat de thesaurus uit 34000 concepten gerelateerd aan kunst en architectuur. De tijdspanne die de AAT beslaat is van de Oudheid tot nu. Elk concept record wordt geïdentificeerd door een unieke numerieke ID. Aan elk concept zijn termen gelinkt, gerelateerde concepten, een parent (voor de plaats binnen de hiërarchie), bronnen voor de data en nota`s. Op deze manier zitten er 131000 termen in de thesaurus.Deze termen worden gebruikt om kunst en architectuur te beschrijven. De termen van een concept bevatten de enkelvoudige vorm, de meervoudsvorm, natuurlijke volgorde , de omgekeerde volgorde, spellingvaranten, verschillende vormen van uitspraak en synoniemen. Onder al deze termen is er één aangeduid als de geprefereerde term of de descriptor.

De AAT is een hiërarchische databank. De root van de hiërarchie wordt de Top van de AAT hiërarchieën genoemd. Buiten de hiërarchische relaties kent de AAT ook nog equivalentie en associatie relaties. Het conceptuele raamwerk van hiërarchieën en facetten is ontworpen om een algemene classificatie te bekomen van de kunst en architectuur. Het raamwerk is niet subject-specifiek. Dit wil zeggen dat er bv. geen termen zijn die specifiek dienen voor het beschijven van een schilderij uit de Renaissance.

De facetten vormen de belangrijkste subdivisies van de AAT hiërachische structuur. Een facet bevat een homogene klas van concepten, waarvan de termen karakteristieken bevatten die hen onderscheidt van termen uit een andere klas. Bv. marmer is een substantie die wordt gebruikt in de creatie van kunst en architectuur en wordt bijgevolg gevonden in het facet over materialen. De facetten zijn copceptueel georganiseerd volgens een schema dat evolueert van abstracte concepten tot meer concrete, fysieke artefacten:

- Geassocieerde concepten: Dit facet bevat abstracte concepten en fenomenen die gerelateerd zijn aan de studie en uitvoering van een brede waaier aan menselijke ideeën en activiteiten, waaronder kunst en architectuur in alle mediavormen, maar ook gerelateerde disciplines. Wat ook wordt gecovered door dit facet zijn de theoretische en critische overwegingen, ideologieën, houdingen en sociale en culturele stromingen. Bv. schoonheid, balans, vrijheid, socialisme, ...

- Fysieke attributen: Dit facet bevat de perceptuele of meetbare eigenschappen van materialen en atrefacten, maar ook eigenschappen van materialen en eigenschappen die niet te onderscheiden zijn als afzonderlijke componenten. Onder deze categorie bevinden zich eigenschappen zoals de grootte en vorm, chemische eigenschappen van materialen, kwaliteiten van textuur en hardheid en eigenschappen zoals oppervlakte afwerking en kleur. Bv. rond, broosheid, grenzen, ...

- Stijlen en Perioden: Dit facet bevat algemeen aanvaarde termen om stylistische stromingen en perioden aan te duiden die relevant zijn voor de lunst en architectuur. Bv. Frans, Louis XIV, Xia, Abstracte Expressionist, ...

- Agenten: Dit facet bevat termen om mensen, groepen van mensen en organisaties te benoemen die worden geïdentificeerd door beroep of activiteit, door fysieke of mentale eigenschappen of door sociale rol. Bv. religieuze orden, corporaties, landschapsarchitecten, ...

- Activiteiten: Dit facet verzamelt alle domeinen van inspanningen, fysieke of mentale acties, systematische sequenties van acties, gebruikte methoden en processen aanwezig in bepaalde materialen of objecten. Activiteiten kunnen variëren van leertrajecten tot enkele gebeurtenissen, van mentale taken tot fysieke acties. Bv. archeologie, ontwerpen, analyseren, tentoonstellingen, corrosie, tekenen, ...

- Materialen: Hier gaat het om termen die een fysieke substantie aanduiden, natuurlijke substanties tot synthetische substanties. Dit facet kan variëren van specifieke materialen tot types van materialen ontworpen voor een bepaalde functie, bv. kleurstoffen, en van gronstoffen tot verwerkte producten. Bv. ijzer, klei, plakband, emulgators, ...

- Objecten: Dit is het grootste facet van alle zeven. Het bevat termen voor alle fysieke of zichtbare objecten die levensloos zijn en het gevolg van een menselijke activiteit.

Hiertoebehoren ook eigenschappen van een landschap die de context leveren voor een bepaald bouwwerk. Bv. schilderijen, facades, kathedralen, tuinen, ...

Een voorbeeld van een record met zijn termen wordt hieronder getoond:

Terms:

still lifes (preferred, C,U,D,American English-P)
still life (C,U,AD,American English)
still-life (C,U,UF,American English)
still-lifes (C,U,UF,American English)
still lives (C,U,UF,American English)
nature morte (C,U,UF,French-P)
nature mortes (C,U,UF,French)
natura morta (C,U,UF,Italian-P)
stilleven (C,U,UF,Dutch-P)
stilleben (C,U,UF,German-P)
vie coye (H,U,UF,French) ..... French for "silent life"; this French term was later replaced by "nature morte"
ontbijtje (H,U,UF,Dutch) ..... Dutch for "small breakfast"
vanitas (H,U,UF) ..... term used to refer to such images in the Netherlands in the 17th century
banketje (H,U,UF) ..... Dutch for "little banquet"
bodegones (H,U,UF,Spanish) ..... term initially used in Spain to describe such images, referring to the lower-class inns and eating-places for which they were painted

Een beetje uitleg bij de gebruikte flags:

D = Descriptor

AD= Alternatieve Descriptor

UF = Use For term, duidt een synoniem aan die geen descriptor of alternatieve descriptor is.

C = Current, actuele term

H = Historische term

B = Beide, actueel en historisch

U = Unknown, ongekend

NA = Not Applicable, niet van toepassing

TGN

Deze thesaurus bestaat uit 912 000 records van plaatsen, gaande van de prehistorie tot nu. De structuur van deze thesaurus is heel gelijkaardig aan deze van de AAT. Een record draait in deze databank rond een plaats. Aan elke record zijn de volgende zaken gerelateerd: unieke numerieke ID, namen, de parent (duidt de plaats in de hiërarchie aan in de TGN

thesaurus), geografische coördinaten, nota`s, bronnen voor de data, andere relaties en een plaatstype. Dit plaatstype beschrijft de rol van de plaats. Bv. een bewoonde plaats of staat of hoofdstad van een staat. Alles bij elkaar bestaan er 1 106 000 namen van plaatsen in de thesaurus. Deze namen worden uitgedrukt in het Engels, maar ook in de lokale taal, eventueel in andere talen en de record bevat ook de historische namen van een plaats. Zoals eerder vermeld bezit een record ook de coördinaten van de plaats. Deze coördinaten zijn eerder een referentie en zijn slechts bij benadering juist.

Net als de AAT thesaurus, is deze thesaurus ook hiërarchisch opgebouwd. De root van de hiërarchie is de Top of the TGN hierarchies. Deze root bestaat uit twee facetten: World en Extraterrestrial Places. Het spreekt voor zich dat het meest bevolkte facet World is. Onder dit facet zijn de plaatsen gerangschikt in subdivisies die de huidige politieke en fysieke wereld representeren, alhoewel er ook historische naties en imperia. Buiten de hiërarchische relaties, kent de TGN ook equivalentie en associatie relaties. TGN wordt hiermee een thesaurus, de ISO en NISO standaard volgend.

Example

Top of the TGN hierarchy (hierarchy root)
.... World (facet)
..... Europe (continent)
..... United Kingdom (nation)
..... [*view physical features*]
..... Anguilla (dependent state) [N]
..... Bermuda (dependent state) [N]
..... Bernicia (historic region)
..... British Antarctic Territory (colony)
..... British Indian Ocean Territory (territory)
..... British Virgin Islands (dependent state) [N]
..... Cayman Islands (dependent state) [N]
..... England (country)

Hieronder vindt u een voorbeeld van de verschillende namen van Brussel, opgslagen in een record van Brussel:

Example
Bruxelles (**preferred**, C,V,N,French-P)
Brussel (C,V,N,Dutch-P)
Bruselas (C,O,N)
Brussels (C,O,N,English-P)
Brusselle (C,O,N)
Brüssel (C,O,N,German-P)
Bruxellae (H,O,N,Latin)

Een beetje uitleg bij de gebruikte flags:

Naam Type Flag:

N = Zelfstandig naamwoord

A = Bijvoegelijk naamwoord

B = Beide, zowel zelfstandig als bijvoegelijk naamwoord

Historische Flag:

C = Current, actueel

H = Historisch

B = Beide, zowel historisch als actueel

U = Unknown, ongekend

NA=Not Applicable, niet van toepassing

Lokale Flag:

V = Vernacular, lokaal

O = Other, ander

U = Undetemined, onbepaald

ULAN

De ULAN is een thesaurus met records rond artiesten. Op dit moment zitten er zo`n 120 000 artiesten in de thesaurus. De structuur is heel gelijkaardig aan deze van de AAT of TGN thesaurus. Elke record krijgt een unieke numerieke Id toegewezen, namen, gerelateerde artiesten, bronnen van de data en nota`s. De records beslaan een periode van de Oudheid tot nu. In totaal zijn er zo`n 293 000 namen opgenomen in de thesaurus. Dit kunnen de gewone namen zijn, maar ook pseudoniemen, andere spellingen van de naam, de naam in verschillende talen en namen die zijn gewijzigd doorheen de tijd bv. door een trouw. Eén van de namen krijgt de flag preferred mee.

Alhoewel de stuctuur redelijk vlak is, de ULAN is toch opgebouwd als een hiërarchische databank. De root van de databank wordt de Top of the ULAN hierarchies genoemd. Deze root heeft twee vertakkingen, facetten: Person en Corporate Body.

Ook deze thesaurus voldoet aan de normen van de ISO en NISO norm. Dit wil zeggen dat er naast hiërarchische relaties ook equivalentie en associatie relaties kunnen gelegd worden tussen de verschillende records.

Hieronder wordt een voorbeeld gegeven van de mogelijke namen van Le Corbusier:

Example
Le Corbusier (**preferred**, **display**, V) .... pseudonym adopted in 1920
Corbusier, Le (V)
Corbusier (V)
Jeanneret, Charles Édouard (V)
Jeanneret, Charles Edouard (V)
Charles Edouard Jeanneret (V)
Jeanneret, Charles-Edouard (V)
Jeanneret-Gris, Charles Édouard (V)
Jeanneret-Gris, Charles Edouard (V)
Gris, Charles Edouard Jeanneret- (V)
Jeanneret-Gris, Charles-Edouard (V)
Le Corbusier, Eduard (V)
Le Corbusier-Saugnier (V)
Corb (V)
Corbu (V)

4.5.4 RAMEAU

Rameau staat voor Répertoire d` Autorité-Matière, Encyclopédique, Alphabétique et Unifié. Het is een thesaurus die alle kennisgebieden bestrijkt in de vorm van een lijst van alfabetische instanties. De ontwikkeling van deze thesaurus is gestart in 1980. Simultaan, maar onafhankelijk werd ook de “Directory of Subject Headings” van de Laval universiteit van Quebec ontworpen, die een vertaling is van de Library of Congress Subject Headings. Vanaf 1983 werd de thesaurus ontwikkeld in samenwerking met het Franse ministerie voor Nationale Educatie (Le Ministère de l`education nationale) en met de de publieke informatiebibliotheek (la Bibliothéque publique d`indormation, BPI) onder de naam LAMECH (Liste d`Autorité Matière Encyclopédique, Collective et Hiérarchisée). In 1987 gingen de twee instituten samenwerken voor het beheer en de verspreiding van de thesaurus die de naam RAMEAU kreeg. Initieel was de thesaurus enkel gevuld met data van de Nationale bibliotheek. Later werd RAMEAU verrijkt met data van de universiteitsbibliotheken. Op dit moment is de thesaurus een nationale indexeringstaal geworden en wordt gebruikt door publieke bibliotheken, een aantal onderzoeksbibliotheken en private organisaties.

RAMEAU is een indexeringstaal die is gestructureerd op drie niveaus:

- Terminologisch niveau: gecontroleerde taal
- Semantisch niveau: hiërarchische taal
- Syntax niveau: geprecoördineerde taal

RAMEAU is dus een gecontroleerde taal. Dit slaat op het feit dat er een controle wordt uitgevoerd op de vorm van de woordenschat, de polysemenen de synoniemen. Elk concept heeft een ”vedette”, hetgeen de geprefereerde term voor dat concept is. Deze vedette moet aan een aantal voorwaarden voldoen:

- Er moet een onderscheid gemaakt worden tussen woorden en uitdrukkingen. Dit is in feite een gevolg van het feit dat de thesaurus precoördinatief is.

  Bv. Travail ; Conditions de travail

- Gebruik steeds het Franse woord, behalve afleidingen van een andere taal.

  Bv. Droit d`auteur (en niet Copyright); Westerns

- Gebruik steeds de meervoudsvorm, maar hier zijn uitzonderingen op (bv. abstracte termen).

  Bv. Vêtements ; Conscience

- Als vedette wordt het meest gebruikt woord genomen.

  Bv. Bicyclettes (en niet Vélocipèdes)

Zoals reeds eerder aangegeven controleert RAMEAU de polysemen. Polysemen zijn woorden met dezelfde schrijfwijze, maar met een verschillende betekenis. RAMEAU maakt een onderscheid tussen de homoniemen, daar er per concept maar één vedette mag zijn. Dit wordt opgelost door:

- Toevoeging van een bepaald woord tussen haakjes.

  Bv. Elasticité; Elasticité (économie politique)

- Toevoeging van een adjectief.

  Bv. Analyse documentaire ; Analyse mathématique

- Onderscheid te maken tussen de enkelvoudige vorm en de meervoudige vorm (verschillende betekenis).

  Bv. Religion; Religions

RAMEAU controleert ook op synoniemen om parallelle indexaties te vermijden. Om deze reden werd de “vedette” ingevoerd. De andere termen die naar hetzelfde concept verwijzen worden als “terme exclu (TE)” aangeduid die naar de vedette verwijzen.

Bv. Bicyclettes EP Vélocipèdes

Bv. Vélocipèdes VOIR Bicyclettes

(Bicyclettes = vedette; Vélocipèdes = TE)

De uitgesloten termen (TE) kunnen synoniemen of quasi-synoniemen zijn, maar ook afkortingen, acroniemen of anders geconstrueerde syntaxen zijn.

Bv. Nantes, Edit de (1598) VOIR Edit de Nantes (1598)

Bv. Religion – Histoire VOIR Histoire religieuse

RAMEAU is ook een hiërarchische taal op semantisch niveau. Dit wil zeggen dat de betekenis van de vedette nog verder wordt gepreciseerd door zijn semantische relaties:

- Hiërarchische relaties TG/TS:

  De generische term wordt aangeduid met TG, de meer specifieke term met TS. Men spreekt hier ook over een vertikale categorisatie.

  Bv. Art TS Architecture

  Bv. Architecture TG Art

- Associatieve relaties TA/TA: Deze relatie zorgt voor de horizontale categorisatie.

  Bv. Architecture TA Construction

  Bv. Construction TA Architecture

Op het niveau van de syntax is RAMEAU precoördinatief. Dit wil zeggen dat complexe onderwerpen, die bv. zijn samengesteld uit verschillende vedettes, op zich een vedette worden. RAMEAU zal niet uit een complex onderwerp automatisch de vedettes halen, maar construeert dus een nieuwe vedette op het moment van de indexering. Een complex onderwerp bestaat uit een hoofdonderwerp (TV – tête de vedette) en de subdivisies. Dit zijn elementen die achter de TV kan gezet worden om het onderwerp meer te preciseren of te vervolledigen.

Bv. Tourisme rural – France = TV + Subdivisie

Hierbij is de volgorde heel belangrijk. Een fout voorbeeld zou zijn: France – Tourisme rural

4.5.5 Thesaurus architecture et patrimoine

Dese thesaurus is een systematische oplijsting van franse architectecturale werken en frans erfgoed. Hij omvat 1135 termen over architecturale werken en 2529 termen over erfgoed. De termen over architecturale werken zijn ontleend uit de indexaties van de databank van Mérimée, deze over het erfgoed komen voort uit de indexaties van de databank over frans erfgoed te Palissy. Het erfgoed behelst architecturale elementen, gebrandschilderde ruiten, meubels, muziekinstrumenten, wetenschappelijke instrumenten, industriële machines en boten.

Deze oplijsting is hiërarchisch opgevat. De termen worden eerst volgens functionele categorieën opgedeeld en deze worden dan telkens verfijnd. Deze functionele categorieën zijn b.v. religieus gebruik, begrafenisgebruik of industrieel gebruik. Zo zijn er achttien basiscategorieën. Deze categorieën zijn disjunctief. Dus geen enkele term komt in twee of meer categorieën voor. Indien de semantiek van de term met andere categorieën overlapt, dan wordt de term in de categorie van voorkeur geplaatst en behelpt men zich door middel van een verwijzing “voir aussi”. Ze bevatten bovendien alle nodige verwijzingen, definities en aantekeningen bij het gebruik.

De thesaurus heeft amerikaanse en engelse versies en er bestaat tevens een italiaanse vertaling van. Op deze manier wordt het internationaal, electronisch raadplegen van de thesaurus bevorderd. Er bestaat evenwel geen xml-versie van de thesaurus, wat het gebruik ervan alleen maar zou vergemakkelijken.

4.6 Overzicht

Titel	Dublin Core The Dublin Core Element Set Version 1.1
Auteur	Dublin Core Metadata Initiative
Uitgever	Dublin Core Metadata Initiative
Date	1999
Website	http://dublincore.org/documents/1999/07/02/dces/
Beschrijving	The Dublin Core is een eenvoudig metadata element set met de bedoeling de inzage van elektronische bronnen te vergemakkelijken. Elementen kunnen gegroepeerd worden in elementen die gegevens bevatten over: inhoud - beschrijving, type, relatie, bron, onderwerp, titel; intellectuele eigendom – datum, formaat, taal, identificatiecode. Het gebruik van deze standaard is opgelegd door meerdere regeringen in Europa en verspreid over de wereld

Titel	MPEG-7
Auteur	MPEG
Uitgever	MPEG
Date	2004
Website	http://www.chiariglione.org/mpeg/standards/mpeg-7/mpeg-7.htm
Beschrijving	Standaard voor het beschrijven van multimedia objecten en de inhoud ervan.

Titel	P/META
Auteur	EBU
Uitgever	EBU
Date	1999
Website	http://www.ebu.ch/en/technical/trev/trev_290-hopper.pdf
Beschrijving	Standaard die gebruikt wordt voor de beschrijving van de inhoud van programma's voor de omroep.

Titel	SMEF-DM
Auteur	BBC
Uitgever	BBC
Date	
Website	http://www.bbc.co.uk/guidelines/smef/
Beschrijving	Standaard die gebruikt wordt voor de beschrijving van de inhoud van programma's voor de omroep.

Titel	MARC/MARC21 Machine-Readable Cataloguing
Auteur	Network Development and MARC Standards Office of the Library of Congress
Uitgever	Library of Congress
Date	2002 (update)
Website	http://www.loc.gov/marc/bibliographic/ebcbdhome.html
Beschrijving	Standaard voor het representeren en communiceren van bibliografische informatie in een machinaal leesbare vorm

Titel	FRBR
Auteur	IFLA
Uitgever	IFLA
Date	1998
Website	http://www.ifla.org/VII/s13/frbr/
Beschrijving	Standaard voor het beschrijven van bibliografische records.

Titel	MODS
Auteur	Network Development and MARC Standards Office of the Library of Congress
Uitgever	Library of Congress
Date	2006 (update)
Website	http://www.loc.gov/standards/mods/
Beschrijving	Standaard voor het beschrijven van bibliografische elementen gericht op gebruik binnen het bibliotheekwezen.

Titel	CDWA
Auteur	Art Information Task Force
Uitgever	Art Information Task Force
Date	2006 (update)
Website	http://www.getty.edu/research/conducting_research/standards/cdwa
Beschrijving	Standaard voor het beschrijven van records uit kunstcatalogen.

Titel	CIDOC-CRM
Auteur	CIDOC Documentation Standards Working Group and CIDOC CRM SIG
Uitgever	CIDOC
Date	2006 (update)
Website	http://cidoc.ics.forth.gr/
Beschrijving	Standaard gebruikt voor het beschrijven van concepten over cultureel erfgoed.

Titel	VRA Core
Auteur	Visual Resources Association's Data Standards Committee
Uitgever	Visual Resources Association's Data Standards Committee
Date	2007 (update)
Website	http://www.vraweb.org/projects/vracore4/
Beschrijving	Standaard gebruikt voor het beschrijven van cultureel erfgoed.

Titel	EAD Encoded Archival Description
Auteur	University of California, Berkeley, Library
Uitgever	University of California, Berkeley, Library
Date	2002 (update)
Website	http://www.loc.gov/ead/index.html
Beschrijving	Standaard voor het beschrijven van collecties, vergelijkbaar met de MARC standaarden. Laat toe het detail van de informatie dat wordt weergegeven te bepalen/aan te passen

Titel	ISAD(G)
Auteur	ICA
Uitgever	ICA
Date	1999 (update)
Website	http://www.ica.org/en/node/30000
Beschrijving	Deze standaard moet helpen bij het opstellen van beschrijvingen van collecties en objecten gericht op archiveringsinstellingen.

Titel	ISAAR
Auteur	ICA
Uitgever	ICA
Date	2004 (update)
Website	http://www.ica.org/en/node/30230
Beschrijving	Deze norm verschaft richtlijnen voor het maken van archivistische geautoriseerde beschrijvingen van entiteiten (organisaties, personen en families) betrokken bij de vorming en het beheer van archieven.

Titel	PREMIS
Auteur	Premis Working Group
Uitgever	Library of Congress
Date	2008 (update)
Website	http://www.loc.gov/standards/premis/
Beschrijving	Deze standaard moet helpen bij het opstellen van preservatie metadata met “deep archiving” tot doel.

Titel	FRAR
Auteur	Working Group on Functional Requirements and Numbering of Authority Records (FRANAR)
Uitgever	IFLA
Date	2007 (update)
Website	http://www.ifla.org/VII/d4/wg-franar.htm
Beschrijving	Deze standaard moet helpen bij het opstellen van authority records

Titel	LCSH
Auteur	Library of Congress
Uitgever	Library of Congress
Date	2008 (update)
Website	http://www.loc.gov/aba/
Beschrijving	Thesaurus van indexeringen die wordt onderhouden door Library of Congress.

Titel	GETTY Thesaurri
Auteur	The Getty
Uitgever	The Getty
Date	2008 (update)
Website	http://www.getty.edu/research/conducting_research/vocabularies/
Beschrijving	Thesauri rond kunst en architectuur, artiestenamen en plaatsnamen.

Titel	RAMEAU
Auteur	BnF
Uitgever	BnF
Date	2008 (update)
Website	http://rameau.bnf.fr/informations/produits.htm
Beschrijving	Een thesaurus vergelijkbaar met LCSH, maar een franse versie ervan.

Titel	Thesaurus architecture en patrimoine
Auteur	Les base architecture et patrimoine
Uitgever	Les base architecture et patrimoine
Date	2008 (update)
Website	http://www.culture.gouv.fr/culture/inventai/patrimoine/
Beschrijving	Thesaurus rond het franse erfgoed.

4.7

5 Declaratieve containers

5.1 Inleiding

Dit hoofdstuk behandelt samengestelde informatie objecten en de relaties tussen data en metadata. Dit zijn objecten die beschrijvende, administratieve en/of structurele metadata combineren tot een informatie object. Het voordeel van zo`n objecten is dat ze kunnen worden uitgewisseld en herbruikt en op deze manier ook de interoperabiliteit bevorderen.

METS is een goed voorbeeld van een standaard die zowel beschrijvende metadata, als administratieve metadata en structurele metadata combineert tot een object. Een object binnen METS bevat een beschrijvende metadata sectie (dmdSec), een administratieve metadatasectie (admSec), een sectie die vertelt welke bronnen tot het object behoort (fileSec), een sectie die de hiërarchische structuur weergeeft van het digitale object (structMap), een sectie die zorgt voor het weergeven van hyperlinks tussen de verschillende componenten van een METS-structuur die beschreven zijn in de structMap (structLink) en tot slot een minder gebruikte sectie die de middelen aanlevert om digitale objecten te verbinden met toepassingen of programma code die in combinatie met andere informatie binnen het METS-document worden gebruikt voor het renderen of weergeven van het digitale object (behaviorSec). Deze secties samen beschrijven een METS object, dat op deze manier kan worden uitgewisseld.

Een ander model voor digitale objecten is LOM. LOM gaat zich meer specifiek toeleggen op het beschrijven van leerobjecten. Op deze manier kunnen leerobjecten gemakkelijker worden herbruikt en bevorderen ze het ontdekken van die leerobjecten. Het LOM datamodel specifieert welke aspecten moeten worden beschreven van het leerobject en welke woordenschatten hier in aanmerking voor komen. Verder legt de specificatie ook uit hoe dit model nog verder kan worden uitgebreid. Dit datamodel wordt reeds veel gebruikt en kent al enkele API`s die LOM ondersteunen.

ORE is een model voor het beschrijven van aggregaties. Deze aggregaties zijn informatie eenheden die wanneer ze worden samengesteld een logisch geheel vormen. Die informatie eenheden kunnen op hun beurt weer een aggregatie vormen. Een voorbeeld hiervan is een boek dat een aggregatie is van hoofdstukken. De hoofdstukken op hun beurt vormen een aggregatie van pagina`s. Het doel van dit model is het hergebruik promoten van de samengestelde objecten. De werkwijze waarop ORE dit doet is via een Resource Map wat een geëncodeerde beschrijving is van de aggregatie. De resource map (ReM) beschrijft de aggregatie die een set van bronnen vormt en mogelijk ook de types en de relaties tussen de

bronnen. Door aan zowel de aggregatie als de Resource Map, die de aggregatie beschrijft, een URI toe te kennen, worden dit gewone webbronnen die kunnen worden uitgewisseld.

Tot slot wordt MPEG-21 DIDL besproken in dit hoofdstuk. Zoals reeds aangehaald in §3.6.1.2 worden in het MPEG-21 raamwerk complexe, digitale objecten beschreven in de Digital Item Declaration (DID) met behulp van de Digital Item Declaration Language (DIDL). DIDL introduceert een set abstracte concepten die tesamen een datamodel vormen voor complexe, digitale objecten. Het DIDL datamodel herkent de volgende entiteiten: een container dat een groep van containers of items is, een item dat een groep van items of componenten is, een component dat een groep van bronnen, een bron dat een individuele datastream voorstelt en tot slot secondaire informatie omtrend een container, item, component of bron. De DIDL specificatie voorziet abstracte definities voor elk van deze entiteiten en hun onderlinge relaties. Hoe de data wordt gestructureerd tot een digitaal object is implementatie afhankelijk. Zo kan b.v. een muziekalbum op verschillende manieren worden beschreven met DIDL. Elke song kan een item zijn, maar het album kan ook worden voorgesteld als een enkel item met als componenten de songs. De representatie van een object zal uiteindelijk afhangen van de doelapplicatie die men voor ogen heeft.

In dit hoofdstuk worden de verschillende informatiemodellen, die hier kort werden voorgesteld, diepgaander besproken.

5.2 METS

De Metadata Encoding en Transmission standaard, kortweg METS, is een specificatie voor het beschrijven en uitwisselen van digitale objecten en hun eigenschappen. METS is een open, niet-propriëtaire standaard die werd ontworpen door de bibliotheekgemeenschap.

METS is XML-gebaseerd en biedt de middelen om metadata op te slaan voor zowel het beheren als het uitwisselen van digitale objecten. Door de XML-basis kent METS een hiërarchische structuur en kan het de hiërarchie uitdrukken van digitale objecten. Een METS-document wordt opgebouwd uit verschillende METS-elementen. Deze elementen worden opgebouwd uit meerdere secties.

```
<mets>
    <dmdSec/>
    <amdSec/>
    <fileSec/>
    <structMap/>
    <structLink/>
```

```
    <behaviorSec/>
</mets>
```

Deze secties voorzien mogelijkheden voor het uitdrukken van de verschillende types metadata (zoals administratieve en beschrijvende) en informatie.

De secties dmdSec (Descriptive Metadata Section) en amdSec (Administrative Metadata Section) dienen als een soort wrappers waarin elementen van andere schema's kunnen worden geplugd. Deze wrappers zorgen er dus voor dat METS uitbreidbaar en modulair is. Voor de inhoud van deze wrappers kent METS geen eigen woordenschat en syntax. Deze worden verzorgd door de standaard die binnen de wrappers worden gebruikt. In de praktijk bestaan er reeds extensie-schema's voor bvb. Dublin Core en MARCXML die gebruik maken van deze techniek. De data in deze wrappers hoeft echter niet strikt tekstueel te zijn, ook binaire formaten zoals MARC21 kunnen hierin worden opgeslagen.

Voorbeeld van een dmdSec en een amdSec:

```
<mets:dmdSec ID="DMD1">
    <mets:mdWrap MIMETYPE="text/xml" MDTYPE="MODS">
        <mets:xmlData>
            <mods:mods version="3.1">
            <mods:titleInfo>
                <mods:title>Interview met een oudstrijder</mods:title>
            </mods:titleInfo>
            <mods:name type="personal">
                <mods:namePart>Jan De Smedt</mods:namePart>
            </mods:name>
            <mods:typeOfResource>audio</mods:typeOfResource>
            </mods:mods>
        </mets:xmlData>
    </mets:mdWrap>
</mets:dmdSec>

<mets:amdSec>
    <mets:rightsMD ID="ADMRTS1">
        <mets:mdWrap MDTYPE="OTHER" OTHERMDTYPE="METSRights">
            <mets:xmlData>
                <rts:RightsDeclarationMD RIGHTSCATEGORY="PUBLIC DOMAIN">
                    <rts:Context CONTEXTCLASS="GENERAL PUBLIC">
                        <rts:Constraints CONSTRAINTTYPE="RE-USE">
                            <rts:ConstraintDescription>
                                Het verdelen en/of kopiëren van dit
                                object is enkel toegelaten mits
                                toestemming van de rechthebbenden.
                            </rts:ConstraintDescription>
                        </rts:Constraints>
                    </rts:Context>
```

```
            </rts:RightsDeclarationMD>
         </mets:xmlData>
      </mets:mdWrap>
   </mets:rightsMD>
</mets:amdSec>
```

Na de dmdSec sectie en de amdSec sectie volgt de fileSec sectie. Deze sectie houdt bij welke bestanden tot het beschreven object behoren. Dit kan gebeuren door het digitale object zelf in te voegen of door een link naar dit bestand op te slaan.

```
<mets:fileSec>
   <mets:fileGrp USE="archive image">
      <mets:file ID="epi01m" MIMETYPE="audio/wav" ADMID="TECHWAV01">
         <mets:FLocat xlink:href=http://www.xxxx.com/01.wav
DOCTYPE="URL"/>
      </mets:file>
   </mets:fileGrp>
</mets:fileSec>
```

Na de fileSec sectie volgt de structMap sectie. In de structMap struct wordt de hiërarchische structuur weergegeven van het digitale object. Dit laat toe de opbouw van het digitale object weer te geven. De structMap sectie laat toe meerdere hiërarchische structuren weer te geven per object. Zo kan men bvb. zowel een logische als een fysische hiërarchie beschrijven. Een interview kan bijvoorbeeld in één bestand zijn opgeslagen (fysische hiërarchie) maar meerdere “onderwerpen” bevatten (logische hiërarchie). Het weergeven van de hiërarchie gebeurt met behulp van divisies.

```
<mets:structMap TYPE="physical">
   <mets:div TYPE="book" LABEL="Het leven tijdens WOII" DMDID="DMD1">
      <mets:div TYPE="page" LABEL="Blank page"/>
      <mets:div TYPE="page" LABEL="Page i: Main title page"/>
      <mets:div TYPE="page" LABEL="Page ii: Blank page"/>
      <mets:div TYPE="page" LABEL="Page iii: Title page"/>
   </mets:div>
</mets:structMap>
```

Tot slot is er de structLink-sectie. Deze zorgt voor het weergeven van hyperlinks tussen de verschillende componenten van een METS-structuur die beschreven zijn in de structMap.

Een minder gebruikte sectie is de zogenaamde behaviorSec sectie. Deze voorziet METS van de middelen om digitale objecten te verbinden met toepassingen of programma code die in

combinatie met andere informatie binnen het METS-document worden gebruikt voor het renderen of weergeven van het digitale object.

METS biedt ook verschillende profielen. Deze dienen als hulp voor het creëren van METS-documenten. Profielen bieden hiertoe een beschrijving van een klasse van METS-documenten in voldoende detail. Voor profielen is een schema beschikbaar. Deze profielen helpen ook programmeurs bij het creëren van software voor het gebruik en de processing van METS-documenten. Verder helpen ze ook bij de interoperabiliteit van digitale bibliotheken.

Een profiel bestaat uit een 13-tal componenten gaande van de titel, een abstract over extension schema's tot een voorbeelddocument.

Voordelen:

- Uitbreidbaar en modulair dankzij de wrapper-secties

Nadelen:

- Mogelijke veiligheidsproblemen bij het invoegen van programmacode in de behaviorSec-sectie
- Kleine community en userbase

5.3 LOM

LOM, of Learning Objects Metadata Standard, is een IEEE-standaard ontworpen om zogenaamde leerobjecten te kunnen beschrijven. Dit kan bijvoorbeeld multimedia content zijn, educatieve content, leerobjectieven, enz. Deze standaard is ontworpen met het oog op het verkrijgen van een minimale set attributen die nodig zijn om de leerobjecten te beheren, lokaliseren en evalueren. De standaard ondersteunt onder andere security, privacy en evaluatie.

LOM definieert een basisschema dat de hiërarchie van data-elementen voor leerobjecten definieert. Op het hoogste niveau bestaan er negen categorieën:

- "General" die algemene informatie bevat over het leerobject in zijn geheel
- "Lifecycle" die informatie bevat over het verleden en de huidige staat van een leerobject, samen met wat het leerobject heeft beïnvloedt tijdens zijn evolutie
- "Meta-Metadata" die informatie bevat over de metadata zelf

- “Technical” die informatie bevat over de technische eisen en karakteristieken van het leerobject
- “Educational” die informatie bevat over het educatieve en pedagogische karakter van het leerobject
- ”Rights” die informatie bevat over de intellectuele eigendomsrechten
- “Relation” die de mogelijkheid biedt de relatie met verschillende leerobjecten weer te geven
- “Annotation” die commentaren kan bevatten over het educatieve gebruik van het leerobject en wanneer en door wie deze commentaren zijn toegevoegd
- “Classification” die het leerobject beschrijft in relatie tot een specifiek classificatiesysteem

Voor elk element specificeert LOM een naam, een uitleg, een grootte, een voorbeeldwaarde, een datatype en nog enkele andere basisdetails. Een voorbeeld van zo een element is “Technical.Location”. Dit is een element “Location” binnen het element “Technical”. Dit element geeft informatie over de plaats van het leerobject, bv. een URL. Sommige elementen kennen een beperkte woordenschat. Dit is een lijst van toegelaten waarden. Andere waarden worden evenwel toegelaten, dit ten koste van een lagere semantische interoperabiliteit. LOM laat verder ook toe data-elementen uit te breiden. Deze data-elementen mogen echter geen LOM-elementen vervangen met het oog op semantische interoperabiliteit. Een voorbeeld hiervan is een element “Naam” toevoegen daar dit kan verward worden met het data-element “General.Title”.

Voor LOM zijn er reeds bindingen ontwikkeld naar RDF en XML. Een LOM-element zou er in XML-vorm dan als volgt kunnen uitzien:

```
<lom xmlns="http://ltsc.ieee.org/xsd/LOMv1p0">
  <general>
    <title>
      <string xml:lang="nl">Interview met een oudstrijder</string>
    </title>
    <language>nl</language>
  </general>
  <technical>
    <location type="URI">
      http://www.interviews.org/oudstrijderx3242.mp3
    </location>
  </technical>
</lom>
```

Tot slot voorziet de LOM-standaard ook in een mapping naar Unqualified Dublin Core voorzien.

Voordelen:

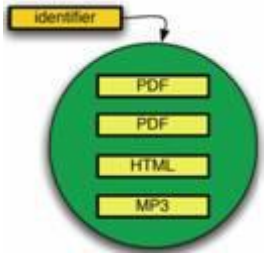
- Heel flexibel en uitbreidbaar
- Mapping voorzien naar Dublin Core en binding met RDF
- Uitgebreid softwareaanbod

Nadelen:

- Geen voorzieningen voor ontologieën

5.4 ORE

Samengestelde informatieobjecten (zie daarom §5 “Compound Objects”) zijn aggregaties van gescheiden informatie-eenheden die een logisch geheel vormen wanneer ze worden samengesteld. Enkele voorbeelden hiervan zijn een gedigitaliseerd boek dat een aggregatie is van hoofdstukken die op hun beurt zijn opgebouwd uit pagina`s. Een ander voorbeeld is een publicatie die een aggregatie is van tekst en ondersteunend materiaal, zoals datasets, software tools, video-opnames van de experimenten, ...



Figuur 5-1: Een samengesteld informatieobject

Verschillende informatiesystemen, zoals content management systemen, leveren ondersteuning voor de opslag en identificatie van samengestelde objecten en de toegang tot de samengestelde objecten en hun geaggregeerde informatie. In de meeste systemen variëren deze componenten volgens semantisch type (artikel, boek, video, dataset, ...), volgens media-type (tekst, beeld, audio, video,...) en volgens mediaformaat (PDF, XML, MP3, ...) of kunnen de componenten op hun beurt weer een samengesteld object zijn. Deze componenten kunnen ook variëren volgens hun netwerklocatie: sommige componenten van een samengesteld object kunnen lokaal opgeslagen zijn, andere op een ander netwerk.

De informatiesystemen verzorgen de opslag en identificatie en leveren de toegang tot deze samengestelde objecten in een architectuurspecifieke manier. Maar omdat het web de facto het platform is voor interoperabiliteit en web-gebaseerde toepassingen zoals search engines geëvolueerd zijn tot de belangrijkste informatiebronnen, zullen deze informatiesystemen hun objecten op het web presenteren. Zij doen dit door een URI te associëren met elke component van het samengesteld object zodat de bronnen door het web via de URI kunnen geïdentificeerd worden. Web services en toepassingen, zoals browsers en crawlers, kunnen deze URI`s gebruiken om de gepaste representaties van de bronnen te verkrijgen.

Jammer genoeg is de manier waarop deze informatiesystemen hun samengestelde objecten publiceren op het web niet perfect en zonder een algemeen aanvaarde standaard. In veel gevallen gaan bepaalde geavanceerde functionaliteiten verloren wanneer deze objecten worden gepubliceerd op het web. Meestal is de publicatie op het web gericht op de eindgebruikers en niet op agents, zoals crawlers. De structuur van het object zit vaak vervat in splash pagina`s, user interface widgets en dergelijke. Deze benadering maakt de essentiële structuur van het object onduidelijk voor machinegebaseerde applicaties zoals browsers, crawlers,...

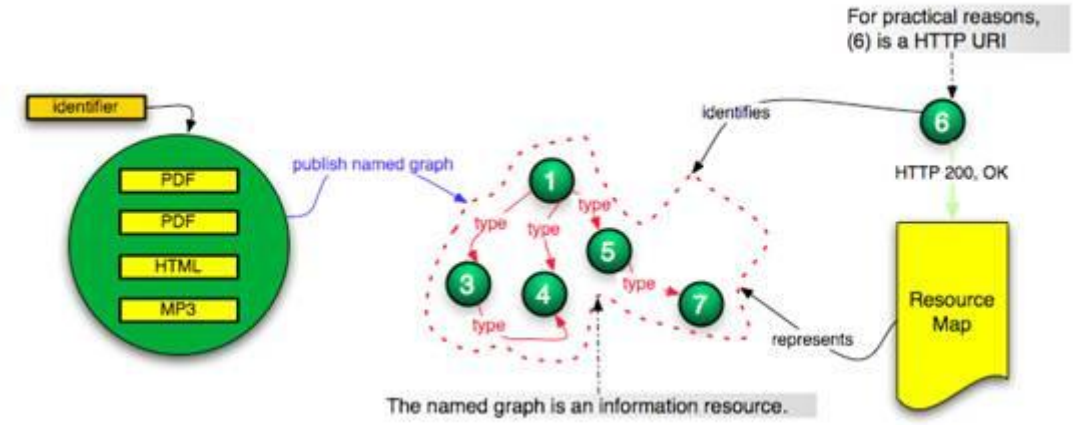
Beschouwen we het voorbeeld weer waar aan alle pagina`s van een gescand boek een HTTP URI is toegewezen. Een webcrawler kan dan op een van die pagina`s landen. De crawler kan vanuit deze pagina weer links vinden naar andere pagina`s van het boek, naar het hoofdstuk dat deze pagina bevat of naar het boek. Naast deze links kunnen er ook links zijn op de pagina naar bijvoorbeeld informatie over de auteur, de uitgever, annotaties,... Een webcrawler kan tussen deze links geen onderscheid maken door het gebrek aan semantiek die in de links vervat zit. De links zijn met andere woorden niet getypeerd of als de links wel type-informatie bevatten, zijn deze niet leesbaar door de webcrawlers. Door de afwezigheid van de standaarden gaat vaak de notie van een samengesteld object met een duidelijke grens en getypeerde relaties tussen zijn componenten verloren.

Het gebrek aan deze standaarden tast de performantie aan van bestaande webservices en toepassingen. Search engines die gebaseerd zijn op crawlers kunnen bruikbaarder worden indien de granulariteit van de resultaten correspondeert met samengestelde objecten in plaats van met de individuele bronnen. De rangschikking van de resultaten van de search engines kan worden verbeterd als de links naar de componenten van een object anders zouden behandeld worden dan de links die verwijzen naar de samengestelde objecten.

Het doel van OAI-ORE (Object Reuse and Exchange) is een gestandaardiseerd, interoperabel en machineleesbaar mechanisme te ontwikkelen die de informatie van de samengestelde objecten kan uitdrukken. De OAI-ORE standaarden zorgen ervoor dat web clients en toepassingen de logische grenzen van de samengestelde objecten en hun relaties onderling kunnen reconstrueren. Dit zal een toegevoegde waarde creëren voor de ontwikkeling van services voor de analyse en de recompositie van samengestelde objecten, zeker in het geval van e-sience en e-scholarship die de doelapplicaties vormen van ORE.

ORE tracht een interoperability layer te realiseren dat een gestandardiseerd middel moet worden om deze repository-specifieke en applicatie-specifieke implementaties van compound objects te publiceren op het web.

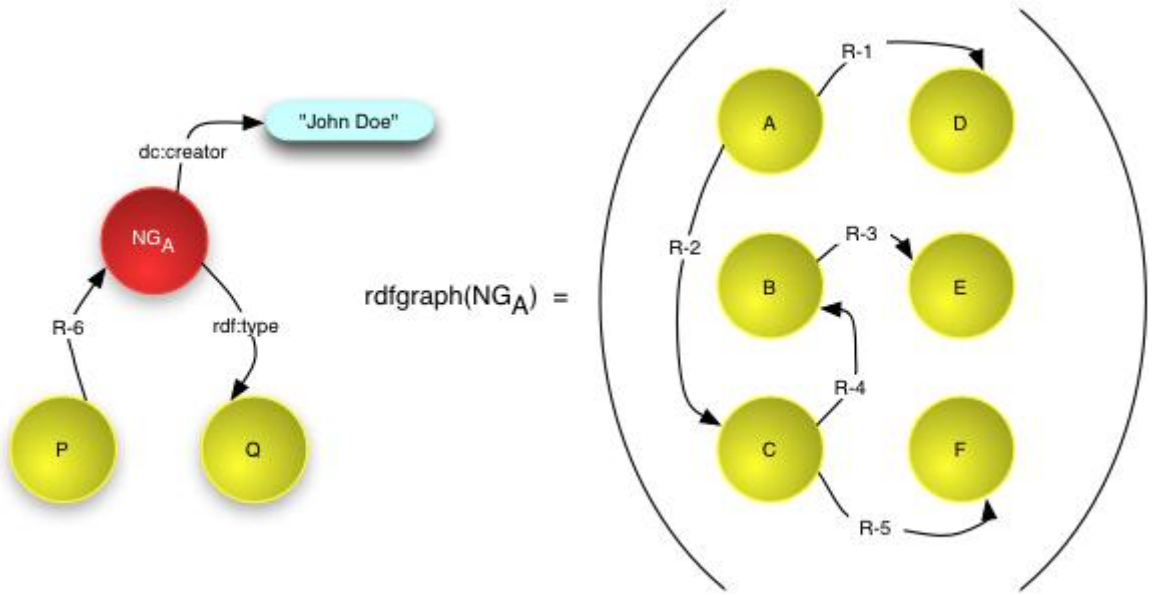
ORE moet dus in staat zijn om de grenzen van compound objects te identificeren. Dit kan gedaan worden door grafen te publiceren op het web die de component/compound object relaties beschrijven. Elke gepubliceerde grafe wordt geidentificeerd door een URI zodat het een gewone webresource wordt. De werkwijze waarop ORE dit doet is via een Resource Map wat een geëncodeerde beschrijving is van de genoemde grafe (named graph).



Een genoemde grafe is een uitbreiding van RDF om een naam te kunnen associëren met een set van triples – een grafe. Zo`n grafe heeft de volgende aspecten:

Een genoemde grafe is een bron, geïdentificeerd door een URI. Deze URI kan zowel subject als object zijn van triples. Dez triples kunnen bijvoorbeeld het type van de grafe aangeven of ze kunnen metadata associëren met de grafe, zoals op de onderstaande grafe is te zien.

De genoemde grafe is niet de RDF-grafe. Het is een bron met een representatie die een set van triples encodeert. De relatie tussen een genoemde grafe en een RDF-grafe die de representatie encodeert, is gedefinieerd via de functie *rdfgraph.*

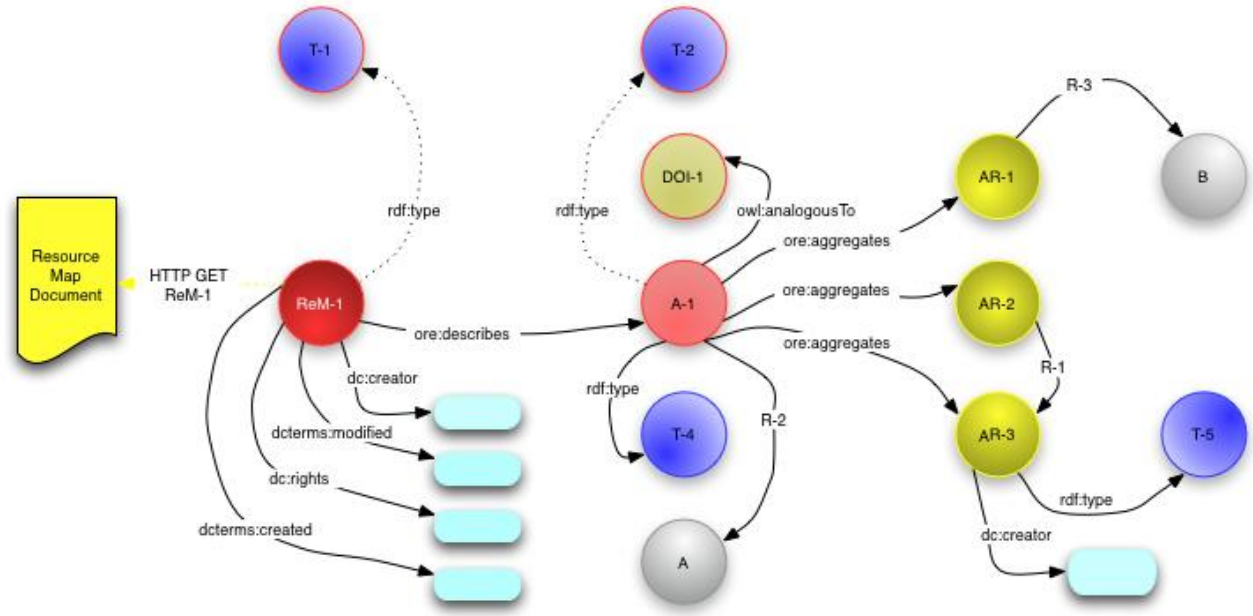


De resource map (ReM) beschrijft een aggregatie die een set van bronnen vormt en mogelijk ook de types en de relaties tussen de bronnen. De bronnen in een aggregatie worden geaggregeerde bronnen genoemd.

Een aggregatie op het web moet een URI (bv. A-1) hebben, wil het via het web bereikbaar zijn. Het ORE-model maakt het noodzakelijk dat een Resource Map precies één aggregatie beschrijft. Een aggregatie kan wel meerdere Resource Maps hebben. Opdat applicaties en clients zouden kunnen refereren naar de aggregatie is het noodzakelijk dat de URI A-1 leidt tot de Resource Map. Dit wordt op de volgende manieren bereikt:

- De URI van de aggregatie kan geconstrueerd worden via een fragment identifier *#aggregation* die wordt toegevoegd aan de URI van de Resource Map.
- Indien in de infrastructuur aggregaties maar één beschrijving mogen hebben, is een andere manier mogelijk: Stel de aggregatie heeft een URI: http://sample.org/ReM-1, dan kan die verwijzen naar de resource map via de URI: http://sample.org/ReM-1.xml of http://sample.org/ReM-1.rdf, afhankelijk van de serialisatie van de ReM.

De onderstaande tekening geeft een volledige representatie ven een aggregatie:



De Rem kan op verschillende manieren worden beschreven. Zo kan men de ReM in RDF/XML of Atom serialiseren.

Voordelen:

- Aggregaties
- Bereikbaar voor webcrawlers
- Binding met RDF
- Hiërarchie

Nadelen:

- Nog in evolutie

5.5 MPEG 21

Het MPEG-21 Multimedia Framework is een open raamwerk dat instaat voor de levering van multimediale data en de definitie van de consumptie ervan en dit voor alle spelers binnen de leverings- en verbruiksketen. MPEG-21 is gebaseerd op twee essentiële concepten: het digitale item, een fundamentele unit voor distributie en transactie, en het concept van de gebruikers, die interageren met deze digitale items. Samenvattend kan men zeggen dat het hoofddoel van MPEG-21 erin bestaat een technologie te definiëren die gebruikers

ondersteunt bij de uitwisseling, de toegang tot, het verbruik, het verhandelen of manipuleren van digitale items.

De gebruiker is een entiteit die binnen de MPEG-21-omgeving interageert met een andere gebruiker of die gebruik maakt van een digitaal item. Gebruikers kunnen individuen, verbruikers, gemeenschappen, organisaties, bedrijven, consortia, regeringen, enz. zijn. Ze worden geïdentificeerd a.d.h.v. hun relatie tot een andere gebruiker voor een bepaalde interactie. Puur technisch maakt MPEG-21 geen onderscheid tussen verbruiker en provider, die dan beide als gebruikers worden beschouwd. Een gebruiker kan op verschillende manieren (publiceren, verbruiken, enz.) van content gebruik maken. Toch kan hij specifieke of zelfs unieke rechten en verantwoordelijkheden hebben, afhankelijk van de interactie met andere gebruikers binnen MPEG-21.

Als basis biedt MPEG-21 een raamwerk waarin een gebruiker over een digitaal item interageert met een andere gebruiker. Een interactie kan onder andere de creatie, archivering of het afleveren van content zijn. MPEG-21 bestaat ondertussen uit 18 delen die men kan onderverdelen in 5 categorieën.

De eerste twee categorieën betreffen ‘declaration’ en ‘identification’. De DID (Digital Item Declaration) is een XML-schema waarin de Digital Item Declaration Language (DIDL) wordt gedefinieerd. Hierin wordt de structuur van complexe digitale objecten beschreven, waaronder de relatie tussen verschillende items. Vandaar dat deze veeleer thuishoren bij de beschrijving van zogenaamde ‘compound objects’.

```
<dia:container>
    <dia:item>
        <dia:component>
            <dia:descriptor/>
            <dia:resource/>
        <dia:component>
    </dia:item>
</dia:container>
```

DII of Digital Item Identification neemt de identificatie van digitale items voor zijn rekening. DII ondersteunt onder andere de unieke identificatie van digitale items en van beschrijvende schema’s en de identificatie van verschillende types digitale items.

```
<Statement>
    <dii:Identifier>
    myID:1234
```

```
    </dii:Identifier>
</Statement>
```

Een derde categorie vormt de DRM of de Digital rights management, die tot doel heeft rechten en toelatingen weer te geven in een machineleesbare vorm. Het uitdrukken van een recht bestaat uit 4 entiteiten en hun wederzijdse relaties: de gebruiker aan wie de rechten zijn toegekend, de rechten zelf, het object waarop de rechten van toepassing zijn en de voorwaarden voor het uitoefenen van de rechten.

Om UMA (Universal Multimedia Access) mogelijk te maken werd verder ook een set normatieve tools ontwikkeld, DIA (Digital Item Adaptation), om een vloeiende adaptatie van digitale items toe te laten. MPEG-21 DIA specifieert aldus de syntax en de semantiek van mogelijke adaptaties. Deze tools kunnen gebruikt worden om resources aan te passen naar gelang de (opgelegde) beperkingen in verband met transmissie, opslag, QoS en/of het afspelen van resources.

De laatste categorie behelst de mogelijkheid om als eindgebruiker te interageren met een digitaal item, DIP (Digital Item Processing). DIP specifieert daarvoor DIM (Digital Item Method) methodes (gebaseerd op een ECMAScipt variant) die binnen een MPEG-21 client applicatie kunnen aangeroepen worden.

6 Digitale archivering: Best Practices.

Zoals eerder aangegeven (§2), kent het OAIS-model als conceptueel raamwerk wijdverbreide toepassingen in verschillende internationale preserveringsprojecten en digitale archiveringssystemen. In het bestek van deze State-of-the-art volstaat een selectief en bondig overzicht van enkele projecten. Vervolgens bespreken we twee praktijkvoorbeelden uit Nederland meer in detail. Die projecten representeren bovendien de twee aspecten waarmee men volgens de OAIS-voorschriften rekening dient te houden. OAIS benadrukt namelijk het onderscheid tussen eisen voor langetermijnbewaring enerzijds en voor consultatie en hergebruik anderzijds. In de ontwikkeling van het e-depot in de KB Den Haag staat langetermijnbewaring centraal en het MultiMatch-project, waar het instituut voor Beeld en Geluid aan deelneemt, concentreert zich in eerste instantie op consultatie en hergebruik.

CASPAR (Cultural Artistic and Scientific knowldege for Preservation access and retrieval) is een project dat mede financierd wordt door de Europese Unie binnen het Sixth Framework Programme (Priority IST-2005-2.5.10, "Access to and preservation of cultural and scientific resources"). Het ging van start op 1 april 2006 en onderzoekt, implementeert en verspreidt innovatieve oplossingen voor digitale preservatie gebaseerd op het OAIS-referentiemodel. 5

Het project **Planets** (Preservation and Long-term Access through Networked Services) situeert zich eveneens binnen het ‘Sixth Framework porgramme en loopt van 2006 tot 2010. Het belangrijkste doel van Planets is de ontwikkeling van diensten en tools die bijdragen tot de langetermijnbewaring van digitale culturele en wetenschappelijke objecten. Het Planets consortium wordt gecoördineerd door de British Library en bestaat verder uit een aantal Europese nationale bibliotheken, archieven, universiteiten en technologiebedrijven. Ook in dit project wordt het OAIS-model expliciet als basismodel aangehaald.6

Het **NDIIP** (National Digital Information Infrastructure and Preservation Program) wordt gecoördineerd door de Library of Congress in samenwerking met instellingen in diverse sectoren zowel binnen als buiten de Verenigde Staten. De beschrijving van de technische infrastructuur van NDIIP is grotendeels gebaseerd op OAIS. 7

Andere voorbeelden van projecten die zich op het OAIS-model baseren:

---

[5] Cf. http://www.casparpreserves.eu/

[6] Adam Farquhar, Helen Hockx-Yu, ‘Planets: Integrated Services for Digital Preservation’, in *International Journal of Digital Curation*, 2 (2007) 2.

[7] http://www.dlib.org/dlib/january07/gladney/01gladney.html

- **Pandora** (Preserving and Accessing networked Documentary Resources of Australia) van de National Library of Australia is een project voor webarchivering.8
- **OCLC Digital Archive** biedt tools voor bibliotheken en archieven met het oog op de archivering van webdocumenten. Daarbij worden enkele onderdelen van het OAIS-model geïmplementeerd, met name Ingest, Store, Disseminate en Administration.9
- **PROV** (Public Record Office Victoria) is een stadsarchief in Australië dat zich baseert op concepten van OAIS.
- **CEDARS** (1998-2002) is een gezamenlijk project van de universiteiten van Oxford, Cambridge en Leeds, dat gericht was op langetermijnbewaring van digitale data. De 'metadata for digital preservation' die daarin voorgeschreven worden, zijn gebaseerd op het OAIS-model.10
- **AIHT** (Archive Ingest and Handling Test) was een onderzoeksproject van de Library of Congress waarbij verschillende universiteitsbibliotheken betrokken waren. Bij het onderzoek naar de efficiëntie van archiefsystemen werd uitgegaan van OAIS-concepten.11

6.1 Ontwikkeling van het E-depot in de KB.

6.1.1 Voorgeschiedenis:

De Koninklijke Bibliotheek van Nederland (KB), gesticht in 1798, is de Nederlandse nationale bibliotheek. Sinds 1974 is de KB ook een depotbibliotheek. In tegenstelling tot andere nationale bibliotheken kent de KB hierbij een vrijwillig depot: uitgevers mogen zelf bepalen of zij publicaties aan de KB schenken. In andere landen is deze depotfunctie van nationale bibliotheken verplicht. In 1993 besliste de KB om haar depottaak uit te breiden met digitale informatie. Zo ontstond de nood aan een systeem om deze digitale publicaties op te slaan en op lange termijn te bewaren.

De KB heeft op het vlak van langetermijnbewaring een pioniersrol vervuld, waarbij ze onder meer talrijke onderzoeksprojecten begeleid heeft. Van 1998 tot 2000 liep onder leiding van de KB het

[8] http://www.nla.gov.au/nla/staffpaper/2003/cathro1.html

[9] Cf. L. Houser, 'OCLC Digital Archive Demonstration', in *Digital Libraries. Proceedings of the 2004 Joint ACM/IEEE Conference on Digital Libraries*, 2004, p. 419. Online: http://ieeexplore.ieee.org/iel5/9280/29473/01336223.pdf ; website OCLC Digital Archive: http://www.oclc.org/digitalarchive/default.htm.

[10] Cf. 'Metadata for Digital Preservation: the Cedars project outline'. Online: http://www.leeds.ac.uk/cedars/colman/metadata/metadataspec.html#note6

[11] M. L. Nelson, J. Bollen, G. Manepalli en R. Haq, 'Archive Ingest and Handling Test. The Old Dominion University Approach', in *D-Lib Magazine*, 11 (2005) 12. Online: http://www.dlib.org/dlib/december05/nelson/12nelson.html.

NEDLIB-project (Networked European Deposit Library), waarin een werkgroep van acht Europese nationale bibliotheken en enkele uitgevers de vereisten van een Europees depotsysteem met betrekking tot langetermijnarchivering onderzocht. Het OIAS-model, op dat moment een ISO-status in wording, werd in het kader van dit project grondig geanalyseerd en voor bibliotheken en archieven verder uitgewerkt.[12]

Het NEDLIB-project heeft veel betekend voor het internationale onderzoek van digitale archivering. De belangrijkste conclusies van het project luidden dat het OAIS-model niet alleen een goed model is voor de archivering van ruimtevaartdata, maar ook een goede basis vormt voor de opzet van digitale archieven in bibliotheken en archieven. In navolging van het OAIS-model concludeerden de NEDLIB-partners dat de functionaliteit voor archivering gescheiden moet worden van de functionaliteit voor zoeken, authenticatie en autorisatie.

Om een e-depot op te zetten volgens de richtlijnen van het NEDLIB-project en het OAIS-referentiemodel bestonden er geen onmiddellijke 'out of the box'-oplossingen. Door middel van een Europese aanbesteding zocht de KB naar een externe technische partner om een elektronisch depot te installeren. In september 2000 tekenden KB en IBM het contract waarmee het DNEP-project ("Depot voor Nederlandse Elektronische Publicaties") werd ingezet. Het DNEP-project bestond uit twee delen. Ten eerste werd nagedacht over de ontwikkeling en implementatie van een groots opgezet digitaal archief, het e-depot. Aangezien de kennis over langetermijnbewaring zich op dat moment in een experimentele fase bevond en er geen definities voorhanden waren van de functionele vereisten voor duurzame bewaring van digitale objecten, omvatte het DNEP-project ook een grondige studie van de noodzakelijke aspecten voor langetermijnbewaring. De resultaten van deze KB/IBM Long-Term Preservation (LTP) Study zijn te lezen op:

http://www.kb.nl/hrd/dd/dd_onderzoek/dnep_ltp_study.html.

De ontwikkeling van het e-depot nam twee jaar in beslag en werd op 12 december 2002, samen met het IBM-systeem DIAS (Digital information and Archival System), als de technische kern van het systeem voorgesteld. Het e-depot is ingericht om Nederlandse elektronische publicaties te bewaren. Daarnaast zal het plaats bieden aan het Nederlands webarchief en masters van gedigitaliseerd materiaal. Omdat de informatievoorziening tegenwoordig een mondiale aangelegenheid is, heeft de KB het e-depot ook opengesteld voor internationale uitgevers als een 'safe space' of 'last resort' voor digitale publicaties.

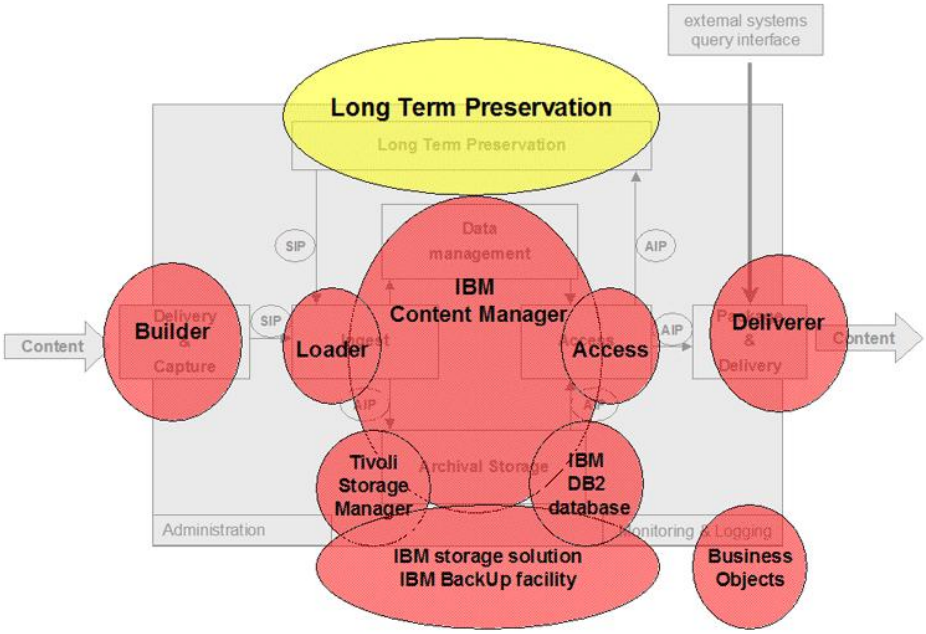
[12] Cf. o.a. T. van der Werf-Davelaar, 'Long-term Preservation of Electronic Publications. The NEDLIB project', in *D-Lib Magazine*, 5 (1999) 9, Online: http://www.dlib.org/dlib/september99/vanderwerf/09vanderwerf.html [verantwoording van de implementatie van OAIS in het Nedlib-project]; C. Lupovici en J. Masanès, 'Metadata for the Long Term Preservation of Electronic Publications' [NEDLIB Report series ; 2], Den Haag: Koninklijke Bibliotheek, 2000.

Ondertussen heeft de KB als een 'safe place' archiveringsovereenkomsten met een groot aantal wetenschappelijke uitgevers, waaronder Elsevier, Kluwer, Biomed Central, Blackwell, Oxford University Press, Springer, Sage Publications, Taylor & Francis (Steenbakkers 2005).

Na de implementatie van het e-Depot in de KB heeft IBM recentelijk nog samengewerkt met de Deutsche Nationalbibliothek voor de implementatie van een digitaal archief "Kooperativer Aufbau eines Langzeitsarchivs digitaler Informationen" (Kopal), gebaseerd op dezelfde technologie als in de KB, met name DIAS (Wollschlaeger 2008).

6.1.2 DIAS-architectuur

De hoofdcomponent van het e-depot is DIAS (Digital Information Archiving System). DIAS was de eerste concrete realisatie van het OAIS-referentiemodel (Steenbakkers 2005).



In figuur 1 wordt duidelijk hoe de componenten van het DIAS-systeem, dat erg modulair opgebouwd is, met de functionele entiteiten van het OAIS-model samenvallen:

- Ingest: de ontvangst en bewerking van de SIP's van de uitgevers
- Archival storage: het opslaan, bewaren en retrieven van AIP's
- Data management: descriptieve info en administratieve data
- Administration: administratie van het dagelijks beheer
- Preservation planning: plannen beheren en uitvoeren van preservatiestrategieën en acties
- Access: toegang tot AIP's en productie van DIP's voor de aflevering aan de Designated Community
- Monitoring en logging: registreren en rapporteren van acties.

In 2003 startte de KB samen met IBM een project voor de ontwikkeling van het 'preservation subsystem' voor DIAS (de gele ovaal in figuur 1). Hiervoor baseerde men zich op de bevindingen van

de onderzoeksrapporten van de LTP-studies in het kader van het DNEP-project over een aantal belangrijke aspecten van digitale bewaring, authenticiteit, mediamigratiemanagement, het archiveren van webpublicaties, e.d.

Volgens de onderzoekers houden langetermijnbewaring en ontsluiting van digitale objecten drie stappen in (Van Diessen and Steenbakkers 2002):

- **Archiveren**:
  het toekennen van een identifier, het digitale object onderbrengen in een gecontroleerde OAIS-compliant-archiefomgeving en de toevoeging van technische en administratieve metadata. De archiefomgeving moet een aparte eenheid in de ICT-infrastructuur zijn. Het is belangrijk dat alle andere functionaliteiten die niet met archivering te maken hebben, zoals zoeken, authenticatie, e.d., apart gehouden worden. Dit verzekert dat de archiefomgeving een duurzaam systeem is dat onafhankelijk van de rest van de ICT-infrastructuur verder kan ontwikkeld worden.
- **De bitstream bewaren**:
  om de bitstream in de originele structuur te bewaren, moeten proactief een aantal stappen ondernomen worden. De bitstream moet op regelmatige basis gekopieerd worden en het medium waarop de bitstream bewaard wordt, moet tijdig 'refreshed' worden.

- **Toegang tot de digitale objecten, ook op lange termijn, garanderen:**
  Om toegang op lange termijn te kunnen garanderen, moeten volgens de DNEP-onderzoeksrapporten de software en de hardware die nodig zijn om het object te kunnen 'afspelen' mee bewaard worden.

Bewaring van digitale objecten op lange termijn moet ten minste vanuit drie perspectieven bekeken worden: bewaring van het medium, technologiepreservatie en intellectuele preservatie.

- Mediumpreservatie heeft te maken met het onderhoud van de dragers waarop informatie is opgeslagen (tapes, diskettes, CD-ROMS). Elke drager heeft een beperkte levensduur. Om te verzekeren dat de data op deze dragers niet verloren gaan, worden de volgende technologische oplossingen voorgesteld: backups, 'refreshing' en checksums om fouten in de bitstream te ontdekken en te verbeteren.
- Snelle veranderingen in de technologie, met name in de bestandsformaten en de software om elektronische informatie te 'renderen', vormen een grotere uitdaging dan mediumpreservatie. Om de opgeslagen informatie ook binnen 50 of 100 jaar te kunnen bekijken of afspelen bestaan er twee oplossingen: migratie en emulatie. Wanneer na verloop van tijd de omvang van het archief toeneemt tot verschillende terabytes aan informatie, zijn dergelijke migratieoperaties geen triviale taak meer. Een migratie van alle objecten in een verouderd bestandsformaat kan dan zodanig lang duren dat de objecten tijdelijk niet beschikbaar zijn.
- Intellectuele preservatie gaat over de integriteit en de authenticiteit van informatie zoals die origineel werd opgeslagen. Authenticiteit betekent dat men het digitale object kan lezen zoals het oorspronkelijk werd opgeslagen. In een digitale omgeving kunnen objecten heel eenvoudig gewijzigd worden. Het is belangrijk dat er technieken of maatregelen ontwikkeld

worden die wijzigingen in de bitstream ontdekken en kunnen verhinderen. Er moet ook een traceerbaar pad bijgehouden worden van het ontstaan van een digitaal object tot de huidige toestand.

Om de authenticiteit van de opgeslagen informatie te bewaren, worden vijf interpretatieniveaus onderscheiden. Elk niveau moet worden beschreven om de authenticiteit van een digitaal object te kunnen bewaren.

- Binary interpretation schemata:
  bepalen hoe fysieke karakteristieken van de hardware (elektrisch, magnetisch, optisch) vertaald worden naar bits en verder gesegmenteerd worden in eenheden van specifieke bitlengte.

- Content interpretation schemata:
  beschrijven specifieke bestandsformaten, die de bits vertalen naar menselijk bruikbare concepten: ASCII, bitmap, …

- Content metatadata interpretation schemata:
  beschrijven hoe bijkomende informatie en karakteristieken geassocieerd worden met bepaalde content data elementen. In het geval van ASCII zijn dit bijvoorbeeld codes voor vette tekst, onderlijnen, enz..

- Structure interpretation schemata:
  beschrijven de relaties tussen de verschillende componenten. Zij bepalen hoe verschillende elementen samengevoegd kunnen worden tot een geaggregeerde eenheid.

- Functional interpretation schemata:
  omvatten de applicatielogica die gebruikt wordt voor het creëren, wijzigen, verkrijgen, verwijderen en renderen van het digital object op een specifieke IT-infrastructuur.

Elk digitaal archief moet bepalen hoe deze authenticiteit bewaard wordt:

1. Beschrijven van de binaire schema's die binnen de IT-infrastructuur gebruikt worden. Bijvoorbeeld 32-bit, 64-bit, bigEndian, lowEndian.
2. Beschrijven van de content schema's voor elk object type. Bijvoorbeeld JPEG2000v1.2, PDF-1992/A.
3. Beschrijven van de metadata content schema's. Bijvoorbeeld Times-New-Roman
4. Beschrijven van de structurele schema's voor elk object. Bijvoorbeeld hoofdstukken, afleveringen, fragmenten.
5. Beschrijven van de functionele schema's en hun impact op de digitale objecten. Bijvoorbeeld om een TIFF-beeld te tonen in een webbrowser moet deze eerst naar JPG geconverteerd worden.

De objectieven en de mogelijkheden van het digitaal archief bepalen welke interpretatieschema's moeten bewaard en geëvalueerd worden.

Op basis van de LTP-studies hebben IBM en de KB een 'preservation subsystem' voor het e-depot ontwikkeld waarin technische metadata geregistreerd worden en waar functionaliteit aan werd toegevoegd die nodig is voor langdurige bewaring. Dit subsysteem bestaat uit drie componenten: de preservation manager voor de registratie van technische metadata, de permanent access toolbox (PATbox) en de preservation processor voor de uitvoer van de preservatieacties zoals migratie en emulatie (Steenbakkers 2005).

In de Preservation Manager wordt informatie geregistreerd over de in het e-depot opgeslagen bestandsformaten. Dit wordt als een essentieel onderdeel van DIAS beschouwd omdat door middel van technische metadata ook toekomstige hardwaresystemen de software bitstream en de bitstream van het digitale object kunnen lezen en gebruikers zo toegang tot de informatie te garanderen.

Dat gebeurt volgens een structuur die van Preservation Layer Models (PLM) en View Paths gebruik maakt. Een Preservation Layer Model beschrijft de verschillende 'lagen' waarop de software draait. Zo kan een PLM bestaan uit de volgende lagen: dataformaat, viewerapplicatie, besturingssysteem en hardwareplatform, waarbij vervolgens deze lagen nader worden gespecificeerd.

Bijvoorbeeld een View Path voor het formaat Pyramid TIFF-formaat kan de volgende elementen bevatten: het platform is Intel Pentium, het besturingssysteem is Windows Vista en de applicatie is Photoshop 7. Telkens wanneer één van deze elementen in onbruik raakt, moet gedacht worden aan migratie of emulatie.

6.1.3 E-depot gegevensarchitectuur

In het kader van WP3 is vooral de gegevensarchitectuur van het e-depot interessant.

Uitgevers bezorgen de elektronische publicatie als een informatiepakket, een SIP, die de volgende onderdelen bevat: essence, metadata, table of contents en eventueel rendering software. De content bitstream wordt opgeslagen in de archiefomgeving. En om duplicatie van metadata te vermijden, werd besloten de descriptieve metadata in de KB-catalogus in een eigen formaat op te slaan. Via de KB-catalogus kunnen eindgebruikers het e-depot raadplegen.

De originele descriptieve metadata en een beperkte set van technische metadata (bestandsformaat, versie van bestandsformaat, grootte van het object,…) zoals die geleverd worden door de uitgevers, worden opgeslagen in de AIP. Specifieke preservatiemetadata worden opgeslagen in de preservation manager.

Nieuwe ontwikkelingen, zowel in de KB als in de digitale bibliotheekwereld, hebben de KB ertoe aangezet na te denken over een vernieuwde gegevensarchitectuur en dit niet enkel voor het e-depot maar voor alle databanken en catalogi in het beheer van de KB.

Eén van de uitgangspunten voor de nieuwe KB-infrastructuur was de vereiste dat de metadata van alle KB-bronnen, inclusief van het e-depot, integraal doorzoekbaar zijn zonder voorkennis van die metadata. Deze integrale doorzoekbaarheid vereist een gemeenschappelijk datamodel voor alle metadata, terwijl in de KB-structuur gebruik wordt gemaakt van verschillende datamodellen voor

verschillende databases. De databases en catalogi die de KB beheert, worden via verschillende websites aangeboden en hebben vaak een eigen metadataformaat, meestal toegespitst op een specifiek materiaaltype of een specifieke functionaliteit. Door het bijzonder karakter van de verschillende metadatastandaarden (bvb. MARC, ISAD, …) en de beschreven objecten, is het moeilijk om één van deze standaarden als gemeenschappelijk datamodel te nemen (Doorenbosch and van Veen 2007).

In plaats van 'proprietary' formaten wilde men gebruik maken van internationale standaarden voor bibliografische en structurele metadata. Specifiek voor langetermijnbewaring in het e-depot onderzocht men de relevantie van de Premis data dictionary voor preservatiemetadata. Bovendien zouden naast wetenschappelijke publicaties nu ook andere materialen opgenomen worden in het e-depot (langetermijnopslag van websites en gedigitaliseerd cultureel en wetenschappelijk erfgoed). Deze ontwikkelingen hebben een belangrijke impact op het e-depot. Zo zal de diversiteit van formaten toenemen en de structuur van de objecten zal complexer worden. Daarom volstond het oude metadatamodel niet meer (Sierman 2007).

Dublin Core was volgens de KB het enige formaat dat "generiek genoeg is om gebruikt te worden voor materiaalbeschrijvingen in de diverse sectoren van cultuur en wetenschap én dat binnen die sectoren voldoende geaccepteerd is" (Doorenbosch and van Veen 2007). Om bij de mapping van de specifieke metadataformaten naar DC niet te veel informatie te verliezen werd gekozen voor Qualified Dublin Core. Om samengestelde objecten op te slaan en te beschrijven en de locatie van de subobjecten vast te leggen werd gekozen voor MPEG 21/DIDL als containerformaat.

De metadata die met de digitale objecten worden aangeleverd, maken integraal deel uit van het digitale object en worden dan ook samen in de langetermijnopslag bewaard. Om het publiek toegang te bieden tot de opgeslagen objecten worden de uitgeversmetadata omgezet naar DC(X) als primair formaat voor descriptieve metadata en ondergebracht in de vernieuwde KB-gegevensinfrastructuur. Via de KB-portal zijn deze metadata doorzoekbaar en kunnen de objecten opgevraagd worden. Aan de beschrijvingen van de objecten worden technische metadata toegevoegd om het gebruikte bestandsformaat eenduidiger te kunnen karakteriseren en om voldoende gegevens voorhanden te hebben om het object op ieder moment door middel van emulatie of migratie te kunnen hergebruiken.

De metadatarecords bestaan uit één of meerdere blokken met verschillende datamodellen:

- Altijd één DC(X) blok voor integrale doorzoekbaarheid
- Optioneel een ander standaardformaat (MARCXML, EAD, …) om de rijkere originele metadata niet te verliezen
- En optioneel nog andere metadatablokken (bijv. Premis voor preservatiedoeleinden).

6.2 Instituut voor beeld en geluid: Multimatch

Het Nederlands Instituut voor Beeld en Geluid (B&G), opgericht in 1997, verzamelt, conserveert en geeft toegang tot het audiovisueel erfgoed dat uit historisch of cultuurhistorisch oogpunt van nationaal belang is. Daarnaast ontwikkelt en verspreidt het instituut kennis op het gebied van audiovisuele archivering, digitalisering en mediageschiedenis. B&G brengt verschillende voormalige archieven, zoals het RTV-Archief Publieke Omroepen, het Filmarchief van de Rijksvoorlichtingsdienst, het Omroepmuseum, Film en Wetenschap, Smalfilmmuseum en verschillende particuliere collecties, samen.

De collecties omvatten meer dan 700.000 uur aan radio, televisie, film en muziek (waarvan een beperkt deel gedigitaliseerd), 2 miljoen foto's en 20.000 voorwerpen. Daarmee is Beeld en Geluid een van de grootste audiovisuele archieven van Europa.

De doelstelling van B&G is vierledig:

- het bedrijfsarchief van de Nederlandse omroepen zijn,
- het audiovisueel cultureel erfgoed van Nederland bewaren, beheren en ontsluiten,
- dit erfgoed ontsluiten voor het grote publiek via een nieuwe 'Media Experience' (interactief media-musuem),
- een kennisinstituut inzake archivering van audiovisueel materiaal zijn.

B&G is opgericht om het duurzaam behoud van het Nederlandse nationale audiovisuele erfgoed te garanderen en het toegankelijk te maken voor zoveel mogelijk gebruikers: professionals, het onderwijs en het grote publiek. Op termijn wil B&G de audiovisuele collectie migreren naar een digitaal archief. Het instituut beschikt nu reeds over een digitale collectie van 10.000 uur.

B&G neemt deel aan verschillende nationale en internationale onderzoeksprojecten met betrekking tot technologieën voor digitale conservering en ontsluiting van audiovisueel materiaal. Eén van die projecten die hier als voorbeeld wordt aangehaald is MultiMatch (http://www.multimatch.eu), waarin men een Europese versie van het 'Geheugen van Nederland' wil realiseren.

6.2.1 MultiMatch

Het MultiMatch-project ambieert de ontwikkeling van een Europese meertalige zoekmachine voor cultureel erfgoedonderzoek. In het project participeren, naast het Instituut voor Beeld & Geluid, tien andere partners waaronder Fratelli Alinari, Biblioteca Virtual Miguel de Cervantes , OCLC PICA, Dublin City University en Universiteit Amsterdam.

De MultiMatch-zoekmachine-engine 'crawled' indexeert culturele erfgoedsites met gedigitaliseerde objecten, zoals bibliotheken, fotoarchieven, musea en audiovisuele archieven. Behalve tekstuele bronnen behandelt het systeem ook beeldmateriaal en videofragmenten. Bovendien moet het systeem minstens vier talen ondersteunen.

De aandacht gaat vooral uit naar de zoekfunctionaliteit van het systeem. De nadruk ligt dan ook bijna uitsluitend op beschrijvende metadatamodellen. Het probleem is dat in de sectoren van cultureel

erfgoed veel verschillende datamodellen in gebruik zijn, wat voor problemen zorgt wanneer men de verschillende collecties wil samenvoegen tot één doorzoekbaar geheel. Niet alleen het gebruik van verschillende metadatamodellen (MARC, ISAD, VRA,…) maar ook het gebruik van verschillende ontologieën, thesauri of gecontroleerde woordenlijsten (LCSH, AAT, …) in de diverse sectoren staan de semantische interoperabiliteit van die collecties in de weg.

In het kader van het MultiMatch-project achtten de onderzoekers het niet realistisch om een nieuw schema te ontwikkelen om dit vervolgens in de verschillende sectoren te introduceren. Er werd veeleer gezocht naar een gemeenschappelijke standaard waar de specifieke modellen aan gemapt kunnen worden. Bij de keuze of ontwikkeling van deze gemeenschappelijke standaard onderscheidden de MultiMatch-onderzoekers een drietal bepalende factoren:

- verwachtingen van de eindgebruikers m.b.t. de zoekfunctionaliteiten,
- de specifieke kenmerken van de gebruikte metadataschema's van de instellingen die data zullen aanleveren. De aparte schema's moeten (semi-)automatisch aan het gemeenschappelijke model gemapt kunnen worden.
- de specifieke kenmerken van de objecten die beschreven moeten worden.

Het onderzoek naar semantische interoperabiliteit van de verschillende collecties van de partners in MultiMatch start met een inventarisatie van een 40-tal metadataschema's, ontologieën en algemene referentiemodellen die gebruikt worden in de culturele erfgoedsector (D2.1: First Analysis of Metadata in the Cultural Heritage Domain). Een eerste vaststelling hierbij was dat in de verschillende erfgoedsectoren gebruik wordt gemaakt van schema's en ontologieën die specifiek gericht zijn op de beschrijving van objecten in die sectoren en met het oog op hun gebruikers. Elk schema was te specifiek om als gemeenschappelijke standaard te gebruiken. Verder bleek uit de inventaris dat, hoewel er onderlinge crosswalks mogelijk zijn tussen veel van de gebruikte medatatastandaarden, interoperabiliteit tussen de verschillende schema's meestal wordt bekomen door te mappen van en naar Dublin Core (DC).

Het probleem is dat DC minder expressief is dan de meer specifieke schema's, waardoor er steeds informatieverlies optreedt bij het mappen van de specifieke schema's naar DC. Dat kan gedeeltelijk opgelost worden door gebruik te maken van de standaarduitbreidingen op DC, zoals Qualified Dublin Core. Een ander alternatief is mappen naar meer expressieve metadatastandaarden zoals MPEG7/21 of naar referentiemodellen zoals FRBR en/of CIDOC CRM.

In het vervolgonderzoek (D.2.2.1) wordt gezocht naar een datamodel dat interoperabiliteit tussen de heterogene collecties moet toelaten. Dit zogenaamde Multimatch-datamodel moest aan een aantal voorwaarden voldoen:

- Het metadataschema moet in XML uitgedrukt worden om de interactie met semantic web techologieën te vergemakkelijken
- De metadataschema's van de aanleverende instellingen moeten (semi-)automatisch gemapt kunnen worden naar het Multimatch-model.
- Het schema moet het object in zijn geheel en volgens relevante subonderdelen kunnen beschrijven. Het moet dus een hiërarchische opstelling kennen.

- Het schema moet gebruik maken van een geïntegreerde en gedeelde ontologie .

DC is de internationaal meest gebruikte standaard voor de beschrijving van objecten in de culturele erfgoedsector. Bovendien kunnen de relevante elementen van om het even welk metadataschema gemapt worden naar de DC-velden, evenwel met verlies van informatie. Voor de doeleinden van het Multimatch-project wordt Dublin Core niet expressief genoeg bevonden. Een meer expressieve standaard zoals MPEG 7 voldeed echter ook niet omdat die in de eerste plaats ontwikkeld is om audiovisuele objecten te beschrijven en dus minder geschikt is om bijvoorbeeld fysieke objecten en hun kenmerken te beschrijven. Een belangrijker nadeel is dat MPEG 7 momenteel nauwelijks gebruikt wordt in de culturele erfgoedsector.

Uiteindelijk wordt geopteerd om een nieuw Multimatch-metadataschema te ontwikkelen, gebaseerd op DCMI-metadataterms. Gedetailleerde documentatie over het Multimatch-metadataschema is beschikbaar op http://www.dcs.shef.ac.uk/%7Ensi/xsd1.1/default.html.

Naast een gemeenschappelijk metadataschema moet ook een gemeenschappelijke semantiek gehanteerd worden bij de invulling van de waarden in de velden van het schema. Dat is mogelijk door het gebruik van relevante standaardthesauri en gecontroleerde woordenlijsten. In het kader van het Multimatch-project werd beslist de Getty-thesauri te gebruiken omdat ze wijd verspreid zijn en door toonaangevende organisaties gebruikt worden. Bovendien bestaan ze in verschillende vormen, waaronder XML.

De drie Getty-thesauri zijn:

- Getty Arts and Architecture Thesaurus (AAT) => artist descriptions
- Getty Unified List of Artist Names (ULAN) => creators names and information
- Getty Thesaurus of Geographic Names (TGN) => geospatial information

Er diende wel nog gezocht te worden naar een consequente manier om deze woordenlijsten uit te breiden met ontbrekende termen. Samenvattend zal men in het MultiMatch-systeem met het onderstaande metadatamodel werken.

- Intern: het MultiMatch-metadataschema dat een uitbreiding is van DCMI Metadataterms.
- Uitwisseling: mapping van het MultiMatch-schema naar DC met de 15 elementen
- Voor verdere interoperabiliteit binnen de erfgoedsector wordt het MultiMatch-schema gemapt naar CIDOC CRM.

7 Conclusies

In dit rapport werd een overzicht gegeven van opslagformaten, metadatastandaarden en containerstandaarden, die de verschillende niveaus representeren waarop men digitale media dient te beschrijven om hun bewaring op lange termijn te garanderen. Op elk niveau situeren zich immers mogelijke gevaren voor dataverlies indien die beschrijvingen niet adequaat en doordacht gebeuren.

Op het laagste niveau is een digitaal bestand opgebouwd uit bits en bytes die op hardwaresystemen opgeslagen zijn. Deze systemen zijn vaak onderhevig aan zogenaamde '*wear-and-tear'*. Vaste schijven en tapes hebben een beperkte levensduur. In de loop van de tijd kunnen digitale bitstreams zich door externe invloeden, zoals corruptie van de dragers, wijzigen. Op dit laagste niveau zijn er hardware- en softwareoplossingen beschikbaar om deze fouten te herstellen. Door verschillende versies van de digitale bestanden op meerdere plekken op aarde op te slaan, kunnen rampscenario's zoals dataverlies door overstromingen, branden en diefstal, worden voorkomen.

Op een hoger niveau vormen vele bytes in de vorm van digitale bestanden een representatie van de opgeslagen data. Bestands- en compressieformaten zoals JPEG en AVI beschrijven de wijze waarop de bits omgevormd kunnen worden tot een interpreteerbare multimediapresentatie zoals beeld, video en geluid. Bestandsformaten zijn echter tijdsgebonden. In de jaren tachtig en vroege jaren negentig was WordPerfectbijvoorbeeld een gangbaar bestandsformaat voor de bewaring van tekstuele data. Tegenwoordig kunnen weinig teksteditors deze bestanden nog openen. Wanneer een bestandsformaat in onbruik geraakt, zijn er voor archieven maar twee opties mogelijk om de opgeslagen informatie te behouden: 1) migratie van het oude bestandstype naar een nieuw formaat (bijvoorbeeld van WordPerfect naar PDF), 2) door software-emulatie een werkbare WordPerfect-lezer "in leven houden". Beide mogelijkheden hebben voor- en nadelen. Door migratie kan informatie verloren gaan en softwarearchivering door middel van emulatie is zeer complex. Om bestandsformaten op lange termijn toegankelijk te houden en een migratie zonder dataverlies mogelijk te maken, is het gebruik van open standaarden noodzakelijk. Bij gesloten bestandsformaten zijn er altijd softwaretools nodig om de data te renderen. Zoals hierboven beschreven, is de archivering van software (en in extremis een werkende IT-infrastructuur) complexer dan het gebruik van open standaarden. Ook bij de keuze van compressieformaten dient men rekening te houden met open en gesloten compressieformaten en -technieken die compressie zonder verlies garanderen.

De bestandsformaten vormen een representatie van de opgeslagen informatie. Bij multimediale data is echter niet alleen de opgeslagen informatie belangrijk maar ook de *look-and-feel* moet behouden blijven. Indien bijvoorbeeld door migraties van bestandsformaten de resolutie of kleurinformatie in beelden verloren gaat, dan betekent dit een verlies aan informatie. Net zoals bij digitalisering van informatie in analoge vorm veel aandacht moet besteed worden aan het behoud van alle aspecten van het originele object, zo zal bij migratie van digitale bestanden een rijke beschrijving van de *look-and-feel* noodzakelijk zijn.

Digitale informatie is ook een conceptueel object dat altijd in een bestaande IT-infrastructuur geïnterpreteerd moet worden.

De authenticiteit van de digitale informatie is aan grotere gevaren onderhevig dan data in analoge vorm. In het laatste geval is het voldoende om alle karakteristieken van het fysieke object te beschrijven, in het eerste geval dienen de gehele ontstaans- en verwerkingsgeschiedenis gearchiveerd te worden.

Op een nog hoger niveau is contextuele informatie onontbeerlijk voor de interpretatie van het digitale bestand. Op lange termijn zullen de producenten van de informatie immers niet meer beschikbaar zijn om de gearchiveerde dataset toe te lichten. Een datacollectie moet van contextuele data vergezeld worden om een volledige beschrijving van de informatie te bieden die, zonder de hulp van externe experts, voor een welomschreven doelpubliek of Designated Community interpreteerbaar blijft..

Op het hoogste niveau wordt een dataset niet enkel door experts maar ook in organisaties en in een tijdsgebonden discours of jargon geproduceerd. Organisatiestructuren kunnen echter wijzigen of verdwijnen en het discours dat eigen is aan een specifieke (productie)context en eindgebruikersgroep met een gemeenschappelijke achtergrondkennis is eveneens tijdsgebonden. Het is dan ook noodzakelijk voldoende informatie mee over te leveren om de data begrijpelijk te houden.

Preservatie van digitale objecten is daarom vanuit minstens drie perspectieven van belang: preservatie van het medium, preservatie van technologie en preservatie van de intellectuele inhoud.

Rekening houdend met die perspectieven is een gelaagd metadatamodel nodig om de data op de drie respectieve niveaus volledig en nauwkeurig te beschrijven:

- Binaire schema's beschrijven de data tot op bitniveau.
- Technische schema's beschrijven op een hoger niveau hoe bytes vertaald worden naar concepten die door mensen geïnterpreteerd kunnen worden, zoals beeld, video en geluid.
- Descriptieve schema's geven een inhoudelijke beschrijving van de data, titels, auteurs, programma'sen dateringen.
- Preservatieschema's beschrijven relaties tussen de databestanden en geven contextuele informatie. De schema's geven technische en administratieve informatie over de ontstaansgeschiedenis van de data en eventuele wijzigingen die ze ondergaan.
- Structurele schema's geven een beschrijving van alle delen van een digitaal object en de relaties tussen de digitale objecten onderling.

In dit rapport werden ook gangbare descriptieve metadataschemas aangehaald die mogelijk door de verschillende projectpartners en instellingen gebruikt worden. Bestandsformaten voor multimedia hebben een zeer breed toepassingsdomein en worden in principe door alle betrokken instellingen geproduceerd. Voor descriptieve standaarden zijn er echter veel onderlinge verschillen. Zo zijn er voor bibliografische beschrijvingen van boeken in de bibliotheeksector andere velden belangrijk dan voor de beschrijving van archiefstukken in de

erfgoedsector. Beschrijvingen van videobestanden in de omroepsector verschillen van beschrijvingen van videokunst in de museumsector.

Voor sommige sectoren kunnen we refereren naar projecten waarin het ontwerp van een gemeenschappelijke (sectorspecifieke) standaard centraal staat of waarin de gebruikte metadataschema's bevraagd en onderzocht werden. IPEA (Innovatief Platform voor Elektronische Archivering)[13] is een IBBT-project, gericht op de omroepsector, waarin P/Meta als generieke descriptieve metadatastandaard voor de betreffende sector gesuggereerd wordt. Deze internationale standaard blijkt namelijk zeer verdienstelijk voor de B2B-uitwisseling van omroepdata (cf. §4.2.2.1). Voor het digitaal archief van BOM, waarbij de meeste omroepen betrokken zijn, zal P/Meta als omroepstandaard dan ook belangrijk zijn. Uit bevragingen van verschillende erfgoedinstellingen voor het project Erfgoed 2.0, een IBBT-project waarin de digitale interactie tussen erfgoedinstellingen in de lijn van Web 2.0 en Library 2.0 beoogd wordt,[14] blijkt onder meer dat in de erfgoedsector een geleidelijk proces van standaardisatie aan de gang is maar dat dit nog lang niet voltooid is. Een suggestie voor een standaard is dan ook noodzakelijk. Hetzelfde geldt voor de musemsector. Ook hier kunnen we verwijzen naar digitale samenwerkingsinitiatieven zoals het project MOVE (Musea Oost-Vlaanderen in Evolutie)[15], waarvoor op termijn een gemeenschappelijke museumstandaard dient afgesproken te worden. Ten slotte halen we hier ook het recent project 'Van Horen zeggen' aan, waarin met verschillende erfgoedinstellingen onderzocht werd hoe men mondelinge bronnen kan bewaren en ontsluiten.[16] Ook hierin werden verschillende formaten, metadatastandaarden en containerformaten met elkaar afgewogen. Deze haalbaarheidsstudie kon zich baseren op bevindingen van het IBBT-project POKUMON (Podiumkunsten Multimediaal Ontsloten), dat zich richtte op de Vlaamse (digitale) archiefwerking van hoofdzakelijk audiovisuele archiefinstellingen.[17] Het spreekt voor zich dat het BOM-project rekening zal houden met bevindingen en conclusies van deze projecten.

Onder meer op basis van de genoemde projectresultaten mogen we besluiten dat het vinden van een grootste gemene deler die de beschrijvingswijze van àlle mogelijke materiaalsoorten dekt, een onhaalbare opgave is. Iedere sector met zijn specifieke materiaalsoorten en data stelt immers afzonderlijke eisen met betrekking tot metadata. Een dergelijke algemene generieke standaard zou tot onnodig veel (meta)dataverlies leiden terwijl het raadzaam is om met het oog op langetermijnbewaring de detaillistische en volledige metadata van de verschillende sectoren mee te archiveren en te bewaren.

---

[13] Cf. http://www.ibbt.be/nl/project/ipea-0

[14] Cf. o.a. B. de Nil en G. Nulens, 'Erfgoed 2.0. Nieuwe wegen voor digitaal erfgoed', in *e-erfgoed*, online: http://www.ibbt.be/files/documents/erfgoed%202-faro-01-03-2008.pdf

[15] Cf. www.museuminzicht.be

[16] Cf. o.m. http://www.faronet.be/blogs/presentatie-onderzoeksresultaten-project-van-horen-zeggen

[17] Cf. http://www.ibbt.be/nl/project/pokumon-0),

Het gelaagd metadatamodel dat we zullen voorstellen, moet dit probleem ondervangen. Er zal namelijk gestreefd worden naar een model dat in zijn uniforme basislaag zo algemeen mogelijk is en in de verfijningslagen meer specifieke metadata bevat die relevant zijn voor de betreffende toepassingsgebieden. Als tussenlaag suggereren we de verschillende sectoren echter het gebruik van een sectorspecifieke metadatastandaard. Samengevat komt het er op neer dat iedere instelling voor haar materiaal uitmaakt welke specifieke metadata van belang zijn en dat ze dus het eigen archiveringssysteem behoudt. Vervolgens past de instelling, afhankelijk van de sector waartoe ze behoort of van het materiaal dat ze bezit, de afgesproken (en door BOM voorgestelde) sectorspecifieke metadatastandaard toe (bijvoorbeeld MARC voor de bibliotheeksector, EAD voor de archiefsector, P/Meta voor de omroepsector, enz.). Ten slotte zullen de sectorspecifieke metadata, die zoals gezegd ook in het gelaagd metadatamodel opgenomen zullen worden, gemapt worden naar een generieke sectoroverschrijdende metadatastandaard die het beheer en de doorzoekbaarheid van het volledige digitale archief zal mogelijk maken. Voor deze generieke laag wordt vaak (Qualified) Dublin Core geopteerd.

Bij de ontwikkeling van een metadatamodel voor de archivering van digitale multimedia moet men dus rekening houden met metadatabeschrijvingen op alle niveaus, van bitlevelbeschrijvingen tot beschrijvingen van de intellectuele inhoud. Omdit te verwezenlijken zijn descriptieve, technische, administratieve, structurele en contextuele metadata nodig. In zijn generieke basislaag zien de beschrijvingen van de uiteenlopende gearchiveerde digitale materiaalsoorten er identiek uit. Op een fijner niveau worden ook alle sector- en materiaalspecifieke metadata bewaard.

8 Bibliografie

MJPEG from http://developer.apple.com/documentation/QuickTime/QTFF/qtff.pdf.
AAC from http://www.apple.com/quicktime/technologies/aac/.
AIFF from http://www.cnpbagwell.com/aiff-c.txt.
CDWA from http://www.getty.edu/research/conducting_research/standards/cdwa/.
CIDOC CRM from http://cidoc.ics.forth.gr/.
Dirac from http://www.bbc.co.uk/rd/projects/dirac/.
DivX from http://www.divx-digest.com/.
Dublin Core from http://dublincore.org/.
EAD from http://www.loc.gov/ead/.
FLAC from http://www.flac.org/.
FRBR from http://www.loc.gov/cds/FRBR.html.
GETTY thesauri from http://www.getty.edu/research/conducting_research/vocabularies/.
GIF from http://www.w3.org/Graphics/GIF/spec-gif89a.txt.
H.264/AVC from http://iphome.hhi.de/suehring/tml/.
ISAD(G) from http://www.ica.org/en/node/30000.
JPEG from http://www.jpeg.org/.
JPEG-LS from http://www.jpeg.org/jpeg/jpegls.html.
JPEG 2000 from http://www.jpeg.org/jpeg2000/.
LCSH from http://www.lib.utah.edu/instruction/handouts/lcsh.html.
MARC from http://www.loc.gov/marc/.
MJPEG from http://tools.ietf.org/html/rfc2435.
MODS from http://www.loc.gov/standards/mods/.
MP3 from http://www.iis.fraunhofer.de/EN/bf/amm/mp3history/mp3history01.jsp.
MPEG from http://www.mpeg.org/.
MPEG-7 from http://www.chiariglione.org/MPEG/standards/mpeg-7/mpeg-7.htm.
MPEG-21 from http://www.chiariglione.org/mpeg/standards/mpeg-21/mpeg-21.htm.
MXF from http://www.midi.org/techspecs/xmf/xmf.php.
Ogg Vorbis from http://www.vorbis.com/.
ORE from http://www.openarchives.org/ore/.
PNG from http://www.w3.org/Graphics/PNG/.
PREMIS from http://www.oclc.org/research/projects/pmwg/.
Theora from http://www.theora.org/.
TTA from http://tta.sourceforge.net/.
VRA Core from http://www.vraweb.org/projects/vracore4/.
WAV from http://ccrma.stanford.edu/courses/422/projects/WaveFormat/.
WMA from http://www.microsoft.com/windows/windowsmedia/default.mspx.
CCSDS (2002). Reference Model for an Open Archival Information System (OAIS), CCSDS.
Doorenbosch, P. and T. van Veen (2007). "Nieuwe gegevensarchitectuur ondersteunt nieuwe diensten." Informatie Professional **4**.
Doorenbosch, P. and T. van Veen (2007). "Nieuwe gegevensarchitectuur ondersteunt nieuwe diensten." Informatie Professional **4**.
Lavoie, B. F. (2004). Technology Watch Report – The Open Archival Information System Reference Model: Introductory Guide. DPC Technology Watch Series Report, OCLC Online Computer Library. **04-01**.
Sierman, B. (2007). Enhancing Our Data Model with PREMIS. DigCCurr 2007.

Steenbakkers, J. F. (2005). "Digital Archiving in the Twenty-First Century." Library Trends **54**: 33-56.
Van Diessen, R. J. and J. F. Steenbakkers (2002). The Long-term Preservation Study of the DNEP Project: An Overview of the Results, IBM Netherlands, Amsterdam, available at: www. kb. nl/kb/sbo/dd/dd_onderzoek/summary_ltpstudy1. html, IBM/KB Long-term Preservation Study Report Series.
Wollschlaeger, T. (2008). "ETD's as pilot materials for long-term preservation efforts in kopal."